International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

Multilingual Conference Audio Transcription, Speaker Diarization And Translation System

Dr.K.Abhirami[1], Mrs.K.Srividhya[2],Ms.S Aswini[3],Ms.B.Bhuvaneshwari[4], Ms.U.Pavithra[5]
[1,2]Associate Professor, Computer Science and Engineering, Kings College of Engineering, Punalkulam,613303, Tamilnadu, India
[3,4,5]UG - Computer Science and Engineering, Kings College of Engineering, Punalkulam,613303, Tamilnadu, India.
***Email id:** abhirami.cse@kingsengg.edu.in[1], srividhya.cse@kingsengg.edu.in[2], aaswini669@gmail.com[3], bhuvanabala0504@gmail.com[4], u.pavithra2309@gmail.com[5]*

Abstract
The increasing prevalence of multilingual conferences and meetings has intensified the need for accurate communication, reliable documentation, and inclusive accessibility across diverse linguistic groups. Conventional approaches based on manual transcription, speaker identification, and translation are inefficient, costly, and susceptible to errors caused by overlapping speech, accents, background noise, and domain-specific terminology. To overcome these limitations, this work presents an AI-driven framework that integrates Automatic Speech Recognition (ASR), speaker diarization, and Neural Machine Translation (NMT). The system automatically converts spoken audio into text, identifies and segments individual speakers, and translates the content into multiple target languages while preserving semantic meaning and contextual accuracy. The resulting speaker-labeled multilingual transcripts support effective knowledge management, decision-making, and post-meeting analysis. By automating end-to-end meeting content processing, the proposed approach enhances efficiency, reduces human effort, improves accuracy, and promotes inclusive participation in academic, corporate, and international communication settings.

***Keywords:** Automatic speech Recognition, speaker Diarization, Neural Machine Translation, Natural Language processing, AI-based system, and Transformer-based models.*

1. Introduction
In today's globalized environment, multilingual conferences and meetings have become increasingly common across academic, corporate, research, and international domains.Organizations and institutions frequently collaborate across geographic and linguistic boundaries, bringing together participants who speak different languages and possess diverse accents and speaking styles. While such interactions promote knowledge exchange and global cooperation, they also introduce significant challenges in effective communication, accurate documentation, and equitable accessibility of meeting content.Accurate capture and preservation of conference discussions are essential for multiple purposes, including real-time understanding, decision-making, compliance, record keeping, and post-event analysis. However, traditional approaches to meeting documentation rely heavily on manual transcription, speaker identification, and human translation. These methods are labor-intensive, time-consuming, costly, and highly susceptible to errors, particularly in complex meeting environments involving overlapping speech, background noise, technical terminology, and multiple speakers. Even skilled professionals often struggle to maintain consistent accuracy under these conditions, making manual processes impractical for large-scale or frequent multilingual events.Recent advancements in artificial intelligence (AI), particularly in speech processing and natural language processing, have enabled the development of automated systems capable of handling complex audio data more efficiently and accurately. Automatic Speech Recognition (ASR) systems powered by deep learning models can now transcribe spoken language into text with high accuracy across multiple

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

languages and accents. These models are trained on large, diverse datasets, allowing them to perform robustly even in challenging acoustic environments. As a result, ASR significantly reduces the dependency on manual transcription while improving processing speed and consistency.In multi-speaker scenarios, accurate identification of individual speakers is critical for clarity, accountability, and structured documentation. Speaker diarization addresses this requirement by determining "who spoke when" within an audio stream. By analyzing audio features such as pitch, tone, and speech patterns, speaker diarization segments the conversation into speaker-specific portions, producing speaker-labeled transcripts. This capability enhances the usability of meeting records by enabling organizations to track individual contributions, generate structured meeting minutes, and perform detailed post-meeting analysis. Beyond transcription and speaker identification, language diversity remains a major barrier to accessibility and inclusivity. Neural Machine Translation (NMT), particularly transformer-based architectures, has emerged as a powerful solution for translating text across multiple languages. Preserving semantic meaning, contextual nuances, and technical terminology. By integrating NMT into the processing pipeline, meeting content can be translated into multiple target languages, ensuring that participants from different linguistic backgrounds can access and understand the information effectively. Motivated by these technological advancements, this project proposes a unified AI-based framework that integrates Automatic Speech Recognition, speaker diarization, and Neural Machine Translation into a single end-to-end system for multilingual conference documentation. The framework processes uploaded audio recordings to automatically generate accurate, speaker-labeled, and multilingual transcripts suitable for documentation, analysis, and knowledge sharing. By automating the entire workflow, the proposed system improves efficiency, enhances accuracy, reduces human error, and increases accessibility. The applications of this framework are broad and impactful. In academic conferences, it enables researchers to access discussions and presentations in their preferred language. In corporate meetings, it supports reliable documentation of strategic decisions and project discussions. In international summits, it ensures transparent and inclusive communication among global stakeholders. In educational contexts, it makes lecture recordings accessible to learners worldwide. Overall, the proposed AI-based approach transforms multilingual conference management by enabling efficient, accurate, and inclusive communication across linguistic and geographic boundaries.

2. The Proposed Design Aims to Achieve the Following Objectives

- Provide automated multilingual speech-to-text transcription from conference audio recordings, enabling accurate documentation of discussions and reducing manual effort in note- taking[1].
- Analyze and segment conference audio using advanced speaker Diarization techniques to identify, label, and distinguish multiple speakers, ensuring clear attribution and structured transcripts.
- Enable seamless Neural Machine Translation across multiple languages, allowing participants from diverse linguistic backgrounds to access and understand conference content without language barriers.
- Generate intelligent summaries and structured reports using AI-based text processing, helping users quickly extract key insights, decisions, and action points from lengthy discussions[2].
- Design an efficient and user-friendly interface for processing, reviewing, and exporting results, ensuring smooth interaction, easy navigation, and multiple export formats for practical use.

3. Literature Survey

- Enhancing Online Dispute Resolution through Natural Language Processing: A Case Study of Kleros, 2025, Alesia Zhuk, proposes NLP-based summarization and analysis to reduce bias and improve efficiency in decentralized online dispute resolution

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

systems.

- Language Identification Based on Multi-scale Feature Recursive Fusion and Adaptive Loss, 2025, Weiwei Li, introduces a multi-scale fusion model that significantly improves language identification accuracy in noisy multilingual speech[3].
- DiCoW: Diarization-Conditioned Whisper for Target Speaker Automatic Speech Recognition, 2024, Matthew Wiesner, enhances multi-speaker ASR by integrating diarization with Whisper, reducing hallucinations and transcription errors.
- Design and Application of Language Translation System Resource Platform on Basis of Artificial Intelligence, 2024, Ruicai Chen and Caiyun Li, presents an AI-based translation platform achieving high accuracy and real-time multilingual translation.
- Design and Application of Language Translation System Resource Platform on Basis of Artificial Intelligence, 2024, Ruicai Chen, emphasizes scalable and secure translation resource platforms using large-scale linguistic data.
- Towards Lifelong Human-Assisted Speaker Diarization, 2023, Meysam Shamsi, proposes a human- in-the-loop diarization system that continuously adapts to new speakers and environments.
- An Experimental Review of Speaker Diarization Methods with Application to Two-Speaker Conversational Telephone Speech Recordings, 2023, Luca Serafini, analyzes diarization techniques and highlights accuracy–computational cost trade-offs.
- The Impact of Stimulus Modalities and Language Proficiency on Bilingual Writing Production, 2024, Zilong Zhong, demonstrates that multimodal inputs significantly enhance bilingual writing performance.
- Voight-Kampff: Generative AI Detection, Multilingual Text Detoxification, and Multi-author Writing Style Analysis, 2023, Janek Bevendorff et al., presents state-of-the-art methods for detecting AI- generated text and authorship analysis.
- Research Priorities in the Field of Multilingualism and Language Education, 2023, Joana Duarte et al., identifies key research gaps emphasizing cultural relevance and fluency in multilingual language technologies[4].

4. Prosposed System Design and Implementaion

The proposed system architecture is a multi-tier pipeline designed to handle the high dimensional challenges of live audio— specifically noise interference, overlapping speakers, and code-switching (language hopping). The framework is partitioned into five logical layers.

4.1. Data Ingestion and Signal Conditioning

The data ingestion layer supports audio file uploads and live WebSocket streaming. Deep Noise Suppression is applied to reduce background noise and overlapping speech. Silero Voice Activity Detection separates speech from non-speech segments. This filtering reduces unnecessary ASR computation. All audio is normalized to a 16 kHz mono format for model compatibility. Shows system architecture[5].

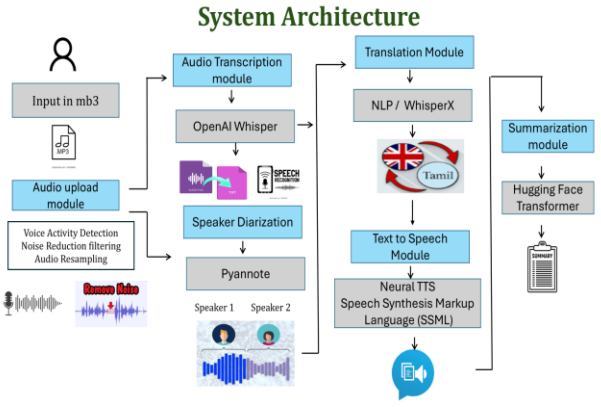

Figure 1 System Architecture

4.2. Neural Identity and Linguistic Logic

This layer maps each utterance to the correct speaker. Pyannote. audio performs speaker diarization using deep speaker embeddings. Agglomerative clustering separates multiple speakers in real time. Known speakers are identified using cosine similarity against

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

stored voiceprints. Language identification using an ECAPA- TDNN model ensures correct linguistic processing. Shows Figure 2 Processing.

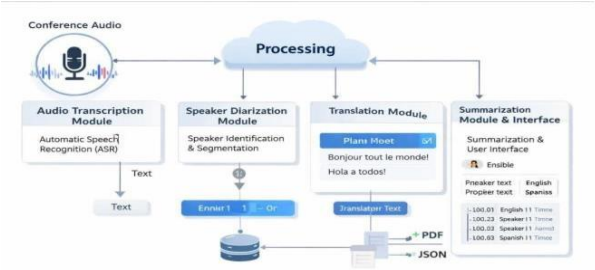

Figure 2 Processing

4.3. Transcription and Temporal Alignment

WhisperX is used as the core speech-to-text engine. VAD-based triggering prevents hallucinations during silent segments. Transcription is performed only on validated speech regions. Wav2Vec 2.0 enables phoneme- level forced alignment. This ensures accurate synchronization between audio and text outputs[6].

4.4. Semantic Post-Processing and Analytics

Raw transcripts are enhanced for readability and structure. A BERT-based model restores punctuation and casing.LLM-based Map-Reduce summarization generates executive summaries[7]. Key decisions and action items are automatically extracted. Sentiment analysis tracks speaker-wise emotional trends over time.

4.5. Interface and Multi-Modal Delivery

The delivery layer presents results to users through an interactive interface. A React-based dashboard displays live transcripts using WebSockets. Each speech segment is tagged with speaker identity and language indicators. Processed data is stored in a structured JSON format. Outputs are exported as PDF reports or SRT subtitle files[8].

5. Methodology

5.1. Data Ingestion and Signal Conditioning

This phase focuses on acquiring and preprocessing raw audio to ensure high-quality input for downstream AI models. Audio is captured from live streams and prerecorded files such as WAV and MP3 and standardized to a uniform 16 kHz mono-channel format. A Voice Activity Detection module separates speech from silence and non-speech events to improve computational efficiency. Spectral noise suppression techniques are applied to mitigate background noise and ambient interference common in large conference environments

5.2. Neural Identity and Linguistic Logic

This module addresses the “who said what” problem by analyzing acoustic features to identify speakers and spoken languages. Speaker diarization segments continuous audio into distinct speaker turns based on voice characteristics using specialized libraries[9]. Voice embeddings are matched against a local database to associate speech segments with known participant identities. Real-time language identification dynamically loads appropriate translation models without manual configuration.

5.3. Semantic Post-Processing and Translation

After transcription, the text is refined to improve readability and usability. Transformer-based Neural Machine Translation models convert transcripts into multiple target languages while preserving context. Text- to-Speech synthesis provides auditory access to translated content. Automated summarization extracts concise meeting minutes, key decisions, and action items from transcripts[10].

5.4.Validation and Success Metrics

System performance is evaluated using quantitative metrics to ensure reliability and scalability. ASR accuracy is measured using Word Error Rate with a target below 10%. Speaker diarization quality is assessed using Diarization Error Rate, aiming for less than 15%. Translation quality is evaluated using BLEU or COMET scores above 30, while system latency is measured using the Real-Time Factor, which must remain below 0.5 to ensure real-time operation[11].

6. Algorithms

6.1. Forced Alignment: WhisperX

Standard ASR systems often lack precise temporal alignment between audio and text. WhisperX resolves this by correcting temporal drift using phoneme-level forced alignment. Individual phonemes are mapped to their exact millisecond positions in the audio waveform. This ensures accurate SRT subtitle generation and real- time

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

caption synchronization.

6.2. Speaker Diarization: Pyannote.audio

This module addresses the "who spoke when" challenge in multi-speaker audio. Using Pyannote.audio, the system extracts speaker embeddings from speech segments. Clustering algorithms group similar embeddings and assign speaker labels such as Speaker 1 and Speaker 2. The framework also supports overlapped speech handling through parallel audio streams[12].

6.3. Neural Machine Translation (NMT) and Summarization

After transcription, alignment, and speaker attribution, the text enters the semantic processing phase. Neural Machine Translation (NMT) uses transformer-basedsequence-to-sequence models to translate content into multiple languages while preserving context. Summarization employs large language models (LLMs) with a MapReduce strategy to generate meeting minutes, key decisions, and action items.

7. Module Classification

7.1. Audio Input Module

The Audio Input Module serves as the entry point of the system. Its primary purpose is to accept conference audio inputs from multiple sources. The module supports both live-streamed audio for real-time processing and pre- recorded audio files such as MP3 and WAV formats for post-event transcription, ensuring flexibility in usage scenarios.

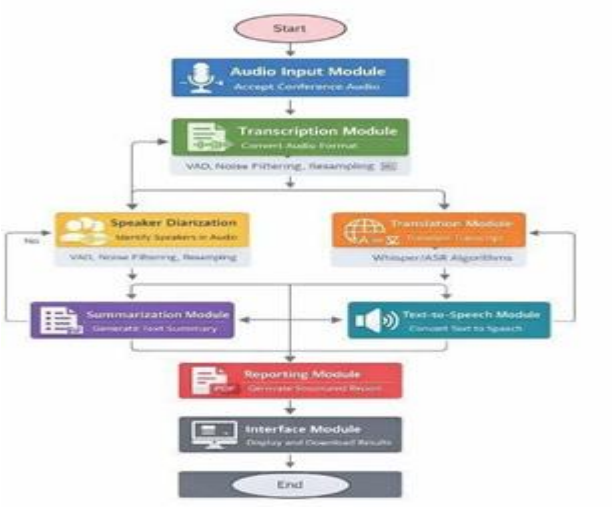

Figure 3 Modules Workflow

7.2. Transcription Module

The Transcription Module is the core engine responsible for converting spoken language into written text. It primarily utilizes OpenAI Whisper, a transformer- based ASR model known for high-fidelity multilingual speech recognition. For optimization, WhisperX is integrated to provide word-level timestamps and forced alignment, ensuring that the generated text precisely matches the original audio timing.

7.3. Speaker Diarization Module

The Speaker Diarization Module addresses the "who spoke when" problem in multi-speaker environments. Its purpose is to identify individual speakers throughout the conversation. Using frameworks such as Pyannote.audio or NVIDIA NeMo, the module extracts voice embeddings (voiceprints) to segment the audio by speaker identity. In the implementation, a Pretrained Speaker Embedding model from SpeechBrain is used to cluster speakers accurately.

7.4. Translation Module

The Translation Module enables cross-lingual communication for international participants. Its main purpose is to translate transcribed text into multiple target languages. The system employs Neural Machine Translation (NMT) based on transformer architectures, ensuring that translations preserve context, semantic meaning, and technical accuracy.

7.5. Text-to-Speech (TTS) Module

The Text-to-Speech (TTS) Module converts translated text back into audio form. This module is especially useful for providing real-time audio interpretation to users who prefer listening over reading captions. It enhances accessibility and supports multilingual auditory output during live conferences.

7.6. Reporting Module

The Reporting Module manages the documentation and archival phase of the system. Its purpose is to export final processed transcripts in professional formats such as PDF, SRT (subtitles), and DOCX. It also integrates summarization to generate meeting minutes and sentiment analysis to provide engagement and interaction metrics.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

7.7. User Interface (UI) Module

The User Interface (UI) Module is the front-facing layer where users interact with system outputs. It displays final results, including live captions, speaker labels, and language indicators. The module typically includes a real-time dashboard and user management tools, enabling smooth interaction and monitoring as the meeting progresses.

8. Implementation Result

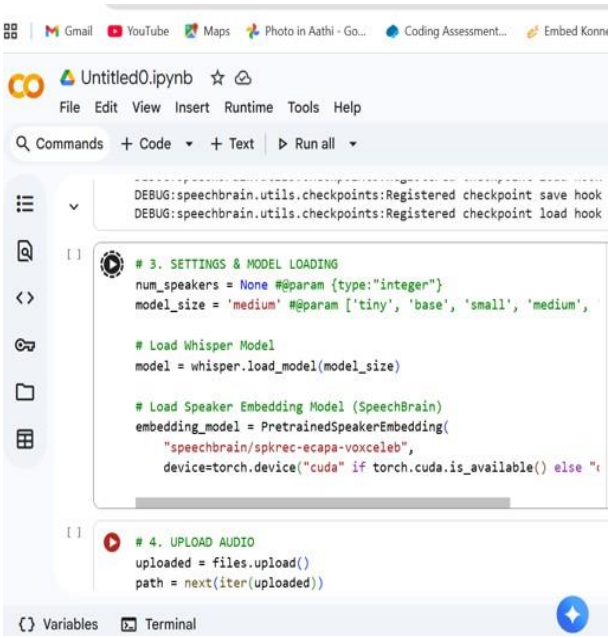

Figure 4 Upload

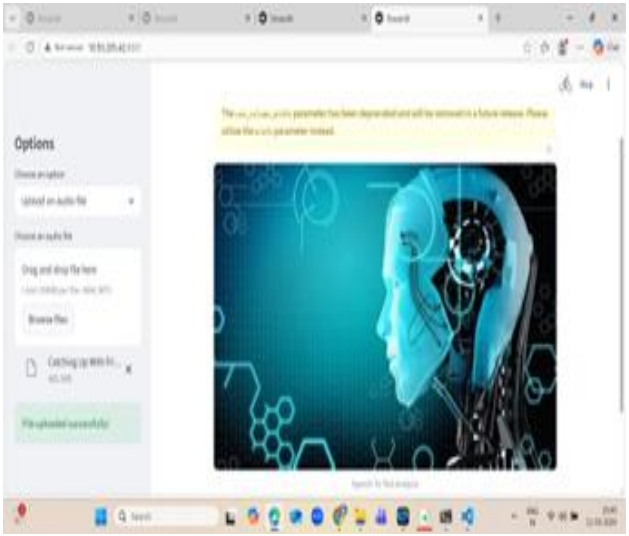

Figure 5 Options

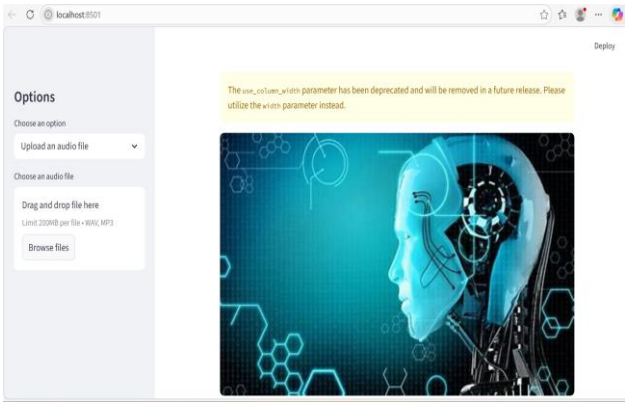

Figure 6 Upload Audio File

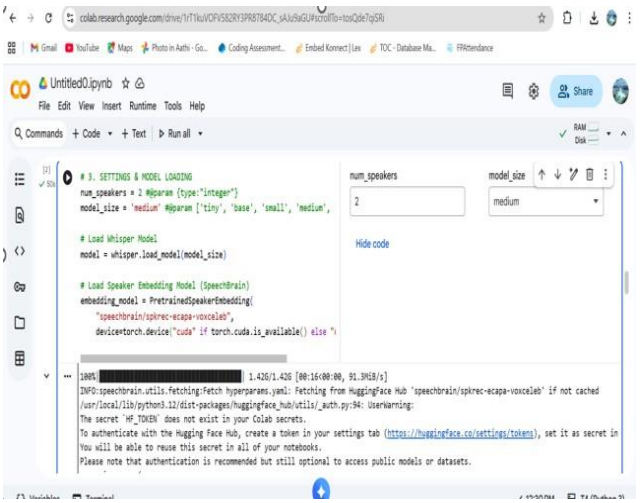

Figure 7 Code

9. Future Work

Live Capabilities: Implementing real-time transcription and translation for live conferences using streaming data pipelines.

Audio Precision: Improving speaker diarization to accurately identify individual voices during overlaps and in noisy environments.

Linguistic Scope: Expanding support for more languages, specifically focusing on regional dialects. Contextual Accuracy: Enhancing translation systems to better recognize and process industry-specific technical terms.

Infrastructure: Making the system scalable for large- scale events and integrating all features into a single, unified platform.

Conclusion

The project successfully demonstrates an AI-based

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1524-1530
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0206

system capable of transcription, speaker identification, and multilingual translation for multi-person conversations. By utilizing the Google Colab environment and models like OpenAI Whisper and Speech Brain, the partial implementation proves that this approach is both feasible and highly beneficial for professional meetings and conferences. While currently a prototype, it establishes a robust foundation for a comprehensive communication tool. Moving forward, the focus will shift toward real-time scalability and contextual precision. Key future developments include

- **Live Integration:** Implementing streaming data pipelines for real-time transcription and translation during live events.
- **Technical & Linguistic Accuracy:** Enhancing diarization for noisy, overlapping environments and improving context-aware translation for technical terminology and regional dialects.
- **Platform Consolidation:** Scaling the infrastructure to support large conferences and integrating all disparate features into one unified platform.Ultimately, this project paves the way for a seamless, inclusive, and automated solution for global communication in any professional setting.

References

[1]. Alesia Zhuk, "Enhancing online dispute resolution through natural language processing: A case study of Kleros," Springer, 2025.
[2]. Weiwei Li, "Language identification based on multi-scale feature recursive fusion and adaptive loss," Springer, 2025.
[3]. Matthew Wiesner, "DiCoW: Diarization-conditioned Whisper for target speaker automatic speech recognition," in Proc. ICASSP, Elsevier,2024.
[4]. Ruicai Chen and Caiyun Li, "Design and application of language translation system resource platform on basis of artificial intelligence," Elsevier, 2024.
[5]. Ruicai Chen, "Design and application of language translation system resource platform on basis of artificial intelligence," Elsevier, 2024.
[6]. Meysam Shamsi, "Towards lifelong human-assisted speaker diarization," Elsevier, 2023.
[7]. Luca Serafini, "An experimental review of speaker diarization methods with application to two- speaker conversational telephone speech recordings," Elsevier, 2023.
[8]. Zilong Zhong, "The impact of stimulus modalities and language proficiency on bilingual writing production," Springer, 2024.
[9]. Janek Bevendorff et al., "Voight-Kampff: Generative AI detection, multilingual text detoxification, multi-author writing style analysis," in CLEF (PAN Lab), 2023.
[10]. Joana Duarte et al., "Research priorities in the field of multilingualism and language education," Review of Educational Research, Springer, 2023.
[11]. A. Bapna, et al., "mSLAM: Massively multilingual joint pre-training for speech and text," arXiv preprint arXiv:2202.01374, 2022.
[12]. Y. Fujita, et al., "End-to-end neural speaker diarization with permutation-free objectives," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2019.