iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2026 **Issue:** Conference **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83194

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com
National Conference on Integrated Approaches to Sustainable Innovation 2026 (IASI 2026)

Towards Privacy-Preserving Speech Intelligence: An Empirical Study of On-Premise Transcription and Diarization in Financial Services

Rahul Kirtikumar Dayal[1], Manish Sharma[2]
Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India

***Abstract*— Voice transcription has become a mission-critical capability for financial services. Regulators require recordings to be retained and want them to be searchable. At the same time, the data inside these recordings are at times are extremely sensitive material, including client PII, trading instructions, advisory conversations, sometimes things the firm itself didn't know it captured. Sending all of that to a public cloud transcription API is looking to have high risk, considering the privacy and sovereignty regulations in each country. Since December 2021, SEC enforcement actions tied to off-channel and record keeping failures have crossed US $3.5 billion.**

This paper takes a practitioner-focused evaluates what it takes to run automatic speech recognition (ASR) and speaker diarization fully on-premise inside a regulated financial institution. This is with a view on the DPDPA 2023, RBI's payment-data localisation circular, and SEBI's cloud framework. The paper describes a five-stage, modular on-premise pipeline built from open-source components: Whisper for ASR, pyannote.audio for diarization, with VAD, post-processing and transcript assembly around them. Leading on-premise options (Whisper, NeMo, wav2vec 2.0, Riva, Vosk, Sortformer) are compared against commercial cloud ASR on accuracy, hardware cost and operational fit. The paper covers real world challenges such as GPU cost, domain vocabulary, hallucinations, the 6-14 month version lag behind closed models.

Keywords*— Data Privacy and Sovereignty, On-premise speech recognition, Automatic speech recognition (ASR), Speaker diarization, Financial services compliance, Voice surveillance, Regulatory technology (RegTech)

I. INTRODUCTION

Over the last few years, Data privacy is getting more and more traction, and regulators are increasingly concerned about data privacy especially related to customer personally identifiable information (PII). Globally, the financial services industry operates in very stringent data protection regulations. One of the key data sets that contain such PII is voice communication. Voice communication such as advisory calls, dealer and trading calls, contact centre calls are recorded and analysed for various purposes. These calls and audio recordings frequently contain PII and proprietary information such as trading strategies. Transcribing such calls is becoming increasing important for regulatory compliance, audit trails, document, improving customer service as well as knowledge management. Various cloud bases public transcription services pose significant risks in regard to data sovereignty, confidentiality and data privacy.

Regulations around the world have imposed strict requirements on processing, storing and transmitting of financial data. Violations of these regulations attract detailed scrutiny and penalties. Since December 2021, the SEC alone has charged penalties exceeding $3.5 billion [14][15]. These actions create a compelling case for on premise AI developments, so that data does not leave the institutions secure environment.

With transformer-based architectures [3], Automatic Speech Recognition (ASR) has gone through a major improvement. Models such as Whisper have been trained with over millions of hours of audio recordings and have managed to achieve Word Error Rates (WER) of less that 3% [4]. Models such as pyannote.audio [5][7] have significantly advanced in diarization managed to achieve diarization error rates of below 10% of certain benchmarks [17]

This paper aims to study the advancement of this technology in line with the regulatory requirements of financial services. It presents a comprehensive study of on-premise voice transcription with speaker diarization, quantifies its importance, evaluates currently available open-source and commercial on premise options with benchmark data, and discuss the challenges and future direction.

©IJRASET: All Rights are Reserved

II. IMPORTANCE OF VOICE TRANSCRIPTION

Voice transcription until recently was a business experiment or a convenience for financial services, however over the last few years its not only become an emerging operational requirement but also a regulatory requirement. There are various use cases where financial services organisations are using Voice Transcription

A. *Regulatory Mandate*

Regulators around the world are mandating recordings of all telephonic or voice conversations especially when it results into a transaction. Different regulators have prescribed timelines for retaining such recordings. In India some regulators require this to be archived for at least 8 years. Financial institutions conduct transactions worth billions of dollars in a matter of minutes, which is done through voice communication. Hence voice transcription and speaker diarization has become a critical regulatory requirement. The financial cost on non-compliance to various regulations is substantial. As shown in Fig. 1, SEC enforcement actions for off channel communication violations have gone up drastically since 2021.

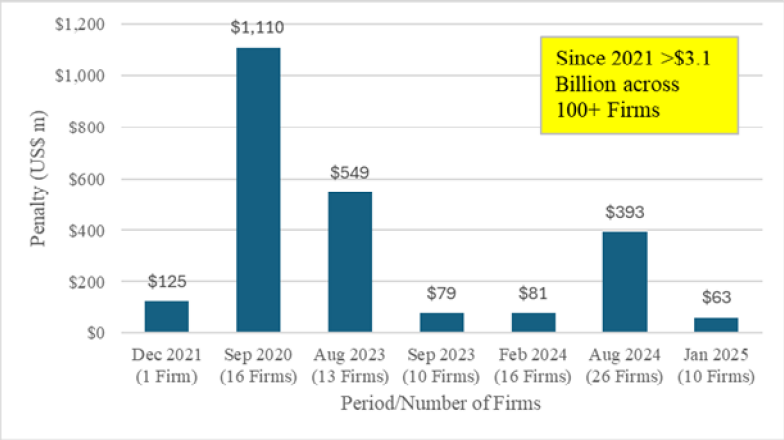


Fig 1: Off Channel Communication & Record keeping Failure Enforcement Actions (Source SEC Press Releases)

B. *Critical Business Use Cases*

Voice transcription and diarization provides several applications with high value benefits especially with Financial Services

1) *Trading Floor and Dealing Room Surveillance*: Trading floors have a large number of transactions worth billions of dollars. Quite often these are money that belong to investors. Real time transcription enables monitoring of these transactions and detection of market abuse, insider trading, front running and such other frauds.
2) *Compliance Monitoring:* Regulators require supervision of conversations of representatives of financial services organisations. This is to ensure there are no pressure selling tactics or unauthorised product recommendations. Audio transcriptions help monitor this and flag any potential misses.
3) *Fraud Detection and AML*: Transcribing of voice calls can help flag any potential fraud or AML cases in the financial services industry, by identifying suspicious words or patterns.
4) *Customer Service:* Real time transcribing of calls can help the contact centre executives understand the sentiment of the customer on call and are able to service them in a much more effective manner. Potential escalations and irate customers can be managed more delicately with the help of real time voice transcriptions.
5) *Contact Centre:* Voice transcriptions of contact centre recording help in quality assurance. Contact centres receive millions of calls a year, and it is practically not possible to listen to all calls for service levels or customer satisfaction. Transcribing these calls and running them past various AI models can help evaluate call quality, service levels and customer satisfaction.

These are only some of the many use cases of voice transcription in financial services. Integration of ASR across financial services workflows provides a significant opportunity for firms to achieve a sustainable competitive advantage.

III. DATA PRIVACY AND REGULATIONS FOR ON-PREMISE IMPLEMENTATIONS

Data Privacy and Data localisation especially of Personally Identifiable Information are gaining more traction. Regulators are issuing various mandates to ensure that customer and investor data is protected.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Voice data quite often contains highly sensitive data such as customer PII, transaction information, investment strategies and advisory conversations. This makes voice data subject to strict regulation, and hence there needs to be control around storage, processing, transmitting and accessing of this data. Globally various regulations such as GDPR, DORA etc, have laid down the principles of data usage and cross border data transfers.

TABLE 1
DATA PRIVACY AND REGULATIONS

Regulation	Jurisdiction	Voice Recording Requirement	Source
MiFID II Art. 16(7)	European Union	All conversations that may result in a transaction	ESMA [12]
FINRA Rule 3170	United States	All phone conversations (taping rule firms)	FINRA [13]
SEC Rule 17a-4	United States	All business communications in WORM storage	SEC [14]
SEC 204-2 (Advisers)	United States	Advice-related oral and written records	Advisers Act 1940
NFA Notice 9070	United States	Futures-related telephone communications	NFA
Dodd-Frank (CFTC)	United States	Oral communications for swap dealers	CFTC
FCA SYSC 10A	United Kingdom	Telephone recording in financial advisory	FCA [19]
EU AI Act 2024/1689	European Union	Transparency, audit, and governance for high-risk AI	EU [1]
DORA 2022/2554	European Union	ICT risk management for financial entities	EU [1]
GDPR	European Union	Voice as biometric / specialcategory data	GDPR
RBI Data Localization	India	All payment system data stored only in India	RBI 2018 [28]
SEBI Cloud Framework	India	All financial data stored within India	SEBI 2023 [29]
IRDAI Guidelines	India	ICT infra and critical data in India	IRDAI [30]
DPDPA 2023	India	Consent-based; SDF localization possible	DPDPA [31]

In India, the regulatory environment has created a strong case for data localisation and sovereignty. The Reserve Bank if India requires that all payments systems data is stored entirely within the boundaries of India. Similarly, The Securities and Exchange Board of India (SEBI) and Insurance Regulatory and Development Authority of India (IRDAI) have mandated data residency and data privacy protection requirements. The Digital Personal Data Protection Act, 2023 (DPDPA) classifies voice recording containing personal data as protected digital data. This means that such data requires customer consent before processing or transferring.
Institutions processing voice data must adhere to all the regulations of data privacy. Voice data processed using external cloud-based AI services carry a huge risk of data exposure, vendor dependency and limited transparency in processing, and could attract stern action by the regulators.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

This creates a strong case for maintaining end to end control over the data process pipelines. On Premise deployment of voice transcription and diarization solutions become a practical and necessary choice. By keeping the data within the organization-controlled infrastructure, institutions can ensure compliance with the regulations and maintain full visibility into the data flows.

IV. VOICE TRANSCRIPTION AND DIARIZATION WITH ON-PREMISE MODELS

This section covers the complete architecture of on Premise Voice Transcription and Diarization built entirely from on premise open-source components. The architecture consists of five stages and follows a modular design philosophy, allowing individual components to be upgraded to new models as and when they are available. All processing occurs on the institutions infrastructure and within the network boundaries, thereby ensuring that no data whether audio/voice data or derived transcripts leave the controlled environment.

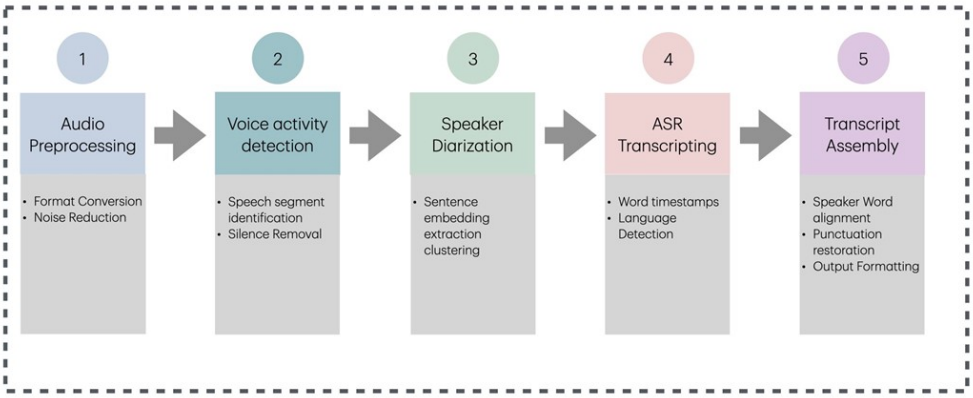


Fig 2: End to End on premise voice diarization pipeline

A. Audio Processing:

Raw audio files from telephony systems, video conferencing platforms or recording devices are in different formats (WAV, MP3, OGG, FLAC) and have diverse sample rates. The first stage is to preprocess and standardise all input files to 16 kHz PCM audio which is the expected format for most modern ASR models [4][5]. Tools such as FFmpeg and python libraries are used for converting these files.

B. Voice Activity Detection (VAD):

VAD identifies regions of the audio that contain speech. The Pyannote audio framework is one such framework that provides a VAD model to provide an output of frame-level speech. NVIDIA offers MarbleNet, which is a light weight model specifically designed for VAD tasks [6]. Pyannote.audio's VAD model has demonstrated higher accuracy rates, especially on challenging audio conditions, however it has higher compute requirements.

C. Speaker Diarization:

This can be a modular approach or an end-to-end approach. Pyannote.audio implements a cascaded approach consisting of 3 phases i. Neural speaker segmentation ii.Speaker embedding extraction iii.Agglomerative clustering that groups embedding into homogeneous clusters.

D. Automatic Speech Recognition (ASR):

This stage is where the speech is recognised. OpenAI's Whisper has emerged as the most dominant open source model for ASR. It has a very low word error rate (WER) which is below 3%, which approaches the human baseline of approx 4%.

E. Transcript Assembly and Post-Processing:

In the final stage, the time stamped diarization output with ASR transcript is generated. It is aligned with speaker segments from the diarization module. The output is a structured transcript with speaker labels, timestamping and context. In the post processing, punctuations are restored, capitalization corrections, numerical formatting and consistency checks are carried out.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. COMPARISON OF ON-PREMISE SOLUTIONS

A wide range of voice transcription and diarization solutions are available with difference license types including open source, freeware and commercial. Each one has their advantages and trade-offs in terms of performance, scalability and ease of deployment.

TABLE II
COMPARISON OF ON-PREMISE SOLUTIONS

Solution	License	ASR WER (clean)	Diarization DER	GPU Req.	Languages
Whisper large-v3	MIT	Yes	N/A	Required	99
Whisper medium	MIT	Yes	N/A	Required	99
pyannote community-1	CC-BY-4.0	N/A	Yes	Recommended	Multi
NVIDIA Sortformer	CC-BY-4.0	N/A	Yes	Required	Multi
NVIDIA NeMo	Apache 2.0	Yes	Yes	Required	Multi
wav2vec 2.0 Large	MIT	Yes	N/A	Required	Multi
Faster Whisper	MIT	Yes	N/A	Optional	99
NVIDIA Riva	Commercial	Yes	Yes	Required	Multi
Vosk (Kaldi)	Apache 2.0	Yes	Basic	CPU OK	20+

1) *Open Source:* OpenAI's Whisper model released under the MIT License is among the most popular and delivers state of the art ASR performance across 99 languages. Pyannote.audio provides the model for speaker segmentation. NVIDIA offers the NeMo model which integrates ASR and diarization under the Apache 2.0 License. Meta offers wav2vec 2.0 that can be fine tuned for domain specific applications.
2) *Freeware and freemium solutions:* In addition to fully open source tools, there are some freeware and freemium solutions such as the community pipeline of pyannote.audio. Hugging face offers free model costing and API interfaces, but it is subject to rate limits.
3) *Commercial on-premise models:* Solutions such as NVIDIA's Riva are optimised ASR and diarization models using TensorRT. Kaldi based platforms from provides such as Speechmatics and Rev.ai, are all commercial models, that have support and customisation capabilities. Some cloud-based vendors have introduced hybrid options; however, they still have partial dependency on cloud. Some solutions such as NICE, Verint and RedBox offer integrated voice capture and transcription systems which are widely used in the financial services industry.

VI. CHALLENGES AND LIMITATIONS

There have been significant advances in on premise deployment of voice transcription and diarization models, however, it still faces some challenges and limitations:

A. *Computational Resource Requirements:*

On premise models have large compute requirements, for example, the Whisper large-v3 requires around 10 GB of GPU memory (VRAM) when running in high precision.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Real-time compliance monitoring systems may require high-performance GPUs such as the NVIDIA A100 or NVIDIA H100 which are priced relatively high and could cost in the range of US $10,000 to US $40,000 per GPU. This requires a large capital outlay for cases where the volume of transcription is high. Further, with the current demand of GPUs there is considerable lead time in delivery of computational resources, which could impact then implementation of such on premise models. While processing call recordings in batches could help reduce compute and hardware requirements, many use cases warrant near real time transcription which will require high compute.

B. Domain Specific and Organization Specific Vocabulary:

Every domain has specialised vocabulary and abbreviations, and Financial services is no different. Financial Instruments, regulatory terminology, and organization specific jargons will not be recognised and interpreted by general purpose models. Models models will require fine-tuning and domain specific training through curated training data.

C. Hallucinations:

AI models are prone to hallucinations and require extensive training to avoid this. Public and commercial models have access to a much larger training data, and feedback loop. Due to this such models get fine-tuned much after and more accurately than on premise models which have restricted data sets and training. On Prem models would require significant training and monitoring to detect any hallucinated transcripts.

D. Version Update Lag:

One of the biggest limitations of on premise deployments and open-source models is the gap in updation and version upgrades compared on commercial public models. Analysis of open vs. closed AI models demonstrates that there is a 6 to 14 month delay between the best best-performing closed commercial models as compared to open-weight models in training and performance [32]. As regards speech specific models, an example of Whispers Large-V3 model was released in November 2023, however, no subsequent major release as of early 2026. In comparison, public models such as Google cloud speech to text, Amazon Transcribe, Azure Speech etc. receive continuous backend updates and improvements. Further, for on premise models, organisations require dedicated ML engineers for model evaluation, regression testing, implementation and rollout - which has cost and time overheads. This is an important consideration while comparing on premise and cloud deployment strategies [32].

VII. FUTURE PROSPECTS

The field of on premise transcription and diarization is progressing rapidly due to its application, and has great prospects for advancement:

A. Unified Models for ASR and Diarization:

Unified models for ASR and speaker diarization will lead to significant simplification. Models such as NVIDIA's Sortformer architecture [9] are moving towards that direction. Future systems are likely to produce diarised output directly from audio input without the need for separate diarization and ASR models.

B. Real-time Streaming Diarization:

Real time streaming diarization is essential for compliance monitoring where regulations require monitoring live conversations such as on trading floors. Models such as NVIDIA's streaming Sortformer are moving in this direction [9].

C. Integration with Large Language Models (LLM):

Integration with On-premise Large Language Models (LLMs) will enable automated summarisation, sentiment analysis, Compliance checking etc. Various open source LLMs such as LLaMA, Mistral, Qwen etc.can be deployed on premise along with the ASR and diarization models to create a more comprehensive intelligent platform.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Federated Learning for Financial Services:

Collaborative training across multiple organisations without sharing raw data could collectively help improve ASR models for domain specific vocabulary, while still maintaining data privacy and data sovereignty. This approach requires industry to work together to improve model accuracy on financial terminology, without sharing the underlying proprietary and PII data.

VIII. CONCLUSIONS

This paper presents a detailed analysis of on-premises voice transcription systems with speaker diarization for financial services. There are three main findings which are presented through this paper. First, voice transcription has become mission critical in financial services due to various use cases, primarily around regulatory frameworks. Non-compliance has resulted in huge fines exceeding US$3.5 billion since only December 2021 [14][15], highlighting the importance of accurate and auditable transcription systems.

Secondly, on premise models have reached performance levels that are close to the commercial cloud models, making them to be practical alternatives that transcribe audio without compromising confidential and PII information. Third, the ASR pipes outlined in this paper enable financial institutions to implement a complete voice transcription and diarization solution onpremise, with full control over confidential, sensitive and PII data. While from a regulatory perspective, these on-premise models meet the data sovereignty requirements, there are several challenges such as high compute requirements, tuning models to domain specific terminology, and managing models in production environments. On going improvements to these models around improved accuracy, real time diarization architecture, federated learning etc are helping reduce these limitations.

Financial Institutions should adopt a phased implementation strategy for on premise models, initially focusing on batch processing of recorded calls and gradually transitioning to near real time transcription as the models and infrastructure improve. The open-source ecosystem provides a strong base allowing institutions to implement on premise AI models that meet the business requirements and also meet the data privacy and regulatory obligations.

IX. ACKNOWLEDGMENT

The authors would like to express their gratitude to the open-source communities behind Whisper, pyannote.audio, NVIDIA NeMo, and the Hugging Face ecosystem, whose contributions have made it possible to have data privacy with AI models for transcription, especially for regulated institutions.

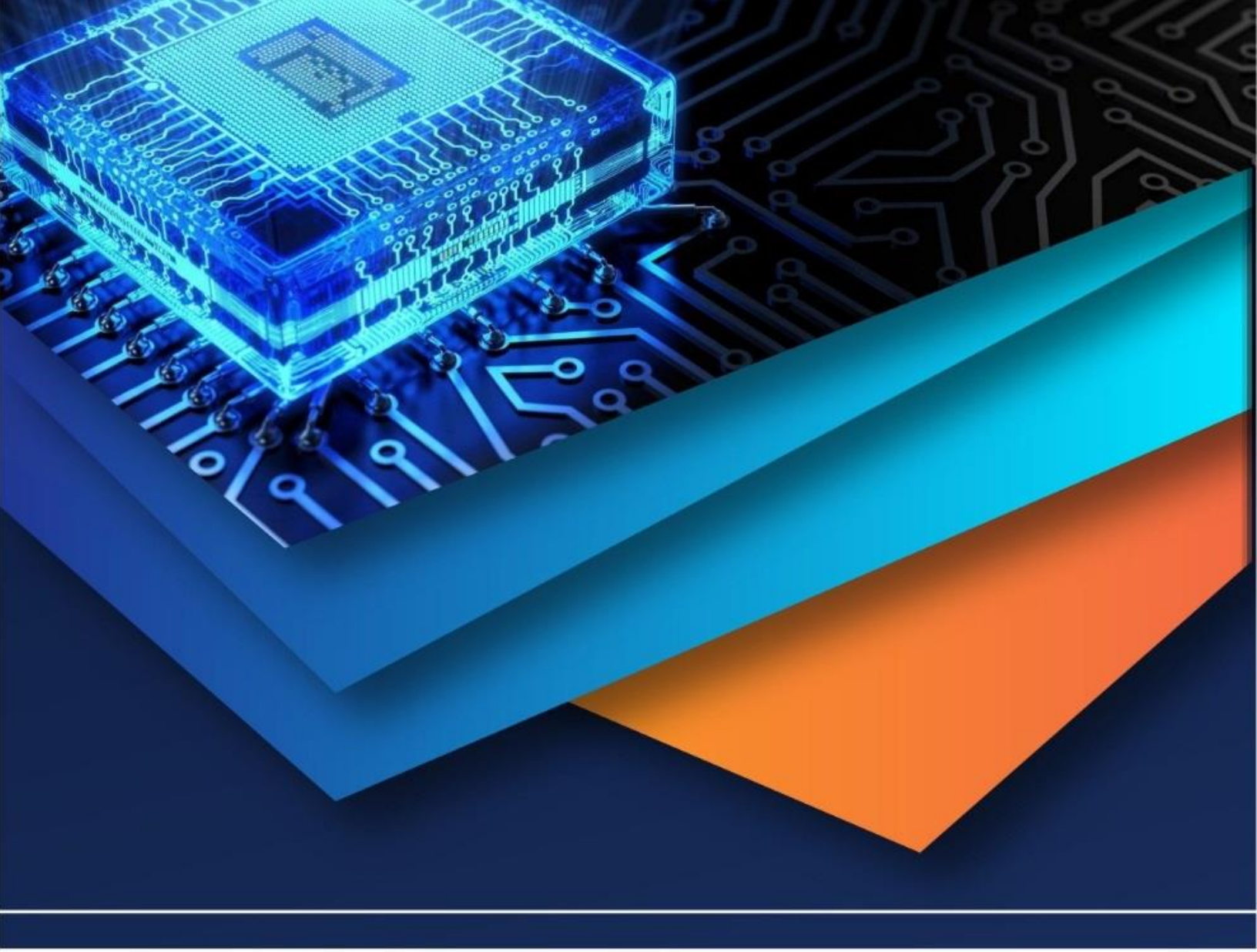
REFERENCES

[1] European Parliament and Council, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (EU AI Act),” Off. J. Eur. Union, Jun. 2024.
[2] U.S. Department of the Treasury, Report on the Uses, Opportunities, and Risks of Artificial Intelligence in Financial Services. Washington, DC, USA, Dec. 2024.
[3] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 5998–6008.
[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), vol. 202, 2023, pp. 28492–28518.
[5] H. Bredin et al., “pyannote.audio: Neural building blocks for speaker diarization,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 7124–7128.
[6] P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of telephone conversations using factor analysis,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2010.
[7] H. Bredin, “pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe,” in Proc. Interspeech, Dublin, Ireland, Aug. 2023, pp. 1983–1987.
[8] J. Yao et al., “Branch-ECAPA-TDNN: A parallel branch architecture to capture local and global features for speaker verification,” in Proc. Interspeech, 2023, pp. 1943–1947.
[9] M. Kunešová, M. Hrúz, Z. Zajíc, and V. Radová, “Detection of overlapping speech for the purposes of speaker diarization,” in Speech and Computer, ser. Lect. Notes Comput. Sci., vol. 11658. Cham, Switzerland: Springer, 2019.
[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: Self-supervised learning of speech representations,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 12449–12460.
[11] H. Bredin and A. Laurent, “End-to-end speaker segmentation for overlap-aware resegmentation,” in Proc. Interspeech, 2021.
[12] European Securities and Markets Authority, “MiFID II Article 16(7) telephone taping requirements,” 2023.
[13] Financial Industry Regulatory Authority, Rule 3170 – Tape Recording of Registered Persons by Certain Firms.
[14] U.S. Securities and Exchange Commission, “SEC charges 16 Wall Street firms with widespread recordkeeping failures,” Press Release 2022-174, Sep. 2022.
[15] U.S. Securities and Exchange Commission, “Sixteen firms to pay more than $81 million,” Press Release 2024-18, Feb. 2024.

©IJRASET: All Rights are Reserved

[16] V. Panayotov et al., "LibriSpeech: An ASR corpus based on public domain audio books," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2015, pp. 5206–5210.
[17] H. Bredin, pyannote.audio: Neural Building Blocks for Speaker Diarization. GitHub repository, Sep. 2025.
[18] M. Kunešová, M. Hrúz, Z. Zajíc, and V. Radová, "Detection of overlapping speech for the purposes of speaker diarization," in Speech and Computer, ser. Lect. Notes Comput. Sci., vol. 11658. Cham, Switzerland: Springer, 2019.
[19] Financial Conduct Authority, SYSC 10A: Recording Telephone Conversations, FCA Handbook.
[20] J. R. Kala, E. Adetiba, A. Abayomi, O. E. Dare, and A. H. Ifijeh, "Speech-to-speech translation with Translatotron: A state-of-the-art review," Results Eng., vol. 28, Art. no. 107780, 2025.
[21] Financial Conduct Authority and Bank of England, AI and Machine Learning in UK Financial Services.
[22] OpenAI, Whisper Large-v3 Model Card. Hugging Face documentation.
[23] OpenAI, Whisper: An Open-Source Speech Recognition System. GitHub repository.
[24] B. Durmuş et al., "SDBench: Benchmark suite for speaker diarization," in Proc. Interspeech, 2025.
[25] New York State Department of Financial Services, Industry Letter on AI Cybersecurity Risks, Oct. 2024.
[26] Y. Pu et al., "Empowering large language models for end-to-end speech translation leveraging synthetic data," in Proc. Interspeech, 2025, pp. 26–30.
[27] S. Zhao et al., "Calm-Whisper: Reduce Whisper hallucination," in Proc. Interspeech, 2025.
[28] Reserve Bank of India, "Storage of payment system data," Circular DPSS.CO.OD No. 2785, Apr. 2018.
[29] Securities and Exchange Board of India, "Framework for adoption of cloud services by SEBI regulated entities," Circular SEBI/HO /ITD/ITD-SEC-1/P/CIR/2023/033, Mar. 2023.
[30] Insurance Regulatory and Development Authority of India, Guidelines on Information and Cyber Security, 2023.
[31] Ministry of Electronics and Information Technology, India, Digital Personal Data Protection Act, 2023 (No. 22 of 2023), Aug. 2023.
[32] Red Hat Developer, "The state of open-source AI models in 2025," Jan. 2026.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)