Overview of the Second Shared Task on Spoken Query Cross-Lingual Information Retrieval for Indic Languages SqCLIR at FIRE 2025

Bhargav Dave[1,*], Prasenjit Majumder[1,2], Debasis Ganguly[3] and Evangelos Kanoulas[4]

[1]*Dhirubhai Ambani University, India*
[2]*GreenAI Services Pvt. Ltd, India*
[3]*University of Glasgow, Scotland, UK*
[4]*University of Amsterdam, Amsterdam, The Netherlands*

Abstract

This paper presents an overview of the Second Shared Task on Spoken Query Cross-Lingual Information Retrieval for Indic Languages (SqCLIR 2025), organised as part of FIRE 2025. The task focuses on developing and evaluating systems capable of retrieving relevant textual documents given spoken queries in Indic languages. Building on the first edition conducted in 2024, SqCLIR 2025 introduced significant enhancements, including the adoption of the IndicMSMARCO dataset as the retrieval collection and the use of spoken queries (audio in .wav format) across five languages — Gujarati, Hindi, Bengali, Kannada, and English. The shared task comprised two subtasks: (1) Spoken Query Ad-Hoc Retrieval (Monolingual), focusing on retrieving documents in the same language as the spoken query; and (2) Spoken Query Cross-Lingual Retrieval, targeting document retrieval across different source and target languages. In addition to the IndicMSMARCO text collection, spoken queries were derived from TREC DL 19 and 20 topics, recorded in Indic languages to simulate realistic voice-search scenarios. System performance was evaluated using standard IR metrics, including nDCG@10, MRR, and recall at multiple depths. A total of six teams registered, though only one team submitted a valid run. Despite limited participation, the task successfully established a foundation for spoken cross-lingual retrieval in low-resource Indic settings, highlighting challenges related to ASR accuracy, language diversity, and speech variability.

Keywords

Spoken Query, Information Retrieval, Indic Language, Cross-Lingual

1. Introduction

Voice-based interfaces have emerged as a dominant means of accessing digital information, driven by the widespread use of smartphones, virtual assistants, and speech-enabled applications. In multilingual societies such as India, users frequently express their information needs through speech in their native languages rather than text. This trend underscores the importance of developing spoken information retrieval systems that can accurately interpret speech queries and retrieve relevant textual content across languages. However, most traditional IR systems are designed for text-based and monolingual settings, which limits their applicability for diverse, multilingual users. Spoken Query Cross-Lingual Information Retrieval (SqCLIR) addresses this gap by combining speech recognition, language translation, and retrieval to enable systems that can process spoken queries in one language and retrieve documents in the same or another language.

Developing effective SqCLIR systems for Indic languages presents several challenges. India's linguistic landscape is characterised by vast diversity, and many languages have rich morphology, distinct scripts, and limited computational resources. The scarcity of parallel corpora and annotated data restricts the development of robust translation and retrieval models. Additionally, Automatic Speech Recognition

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Corresponding author.
bhargavdave1@gmail.com (B. Dave); prasenjit.majumder@gmail.com (P. Majumder); debforit@gmail.com (D. Ganguly); ekanoulas@gmail.com (E. Kanoulas)
0000-0003-2742-480X (B. Dave); 0000-0003-0840-9313 (P. Majumder); 0000-0003-0050-7138 (D. Ganguly); 0000-0002-8312-0694 (E. Kanoulas)

(ASR) systems for Indic languages face high word error rates due to factors such as accent variation, background noise, and frequent code-mixing. These issues make spoken cross-lingual retrieval a complex, multi-stage problem that remains underexplored in low-resource multilingual contexts.

Research in IR has evolved from traditional lexical-matching approaches such as BM25 [1] to dense retrieval models that learn semantic representations using dual-encoder architectures like DPR [2] and ColBERT [3]. These models have significantly improved multilingual and cross-lingual retrieval when combined with language-agnostic encoders such as LaBSE [4]. Work on Cross-Lingual Information Retrieval (CLIR) has been extensively studied in evaluation forums such as TREC, CLEF, and FIRE, where approaches based on query translation, document translation, or shared multilingual embeddings have been compared [5, 6]. In parallel, Spoken Information Retrieval (SIR) and Spoken Document Retrieval (SDR) have investigated the retrieval of textual or audio content from speech inputs [7], although most prior studies relied on ASR transcripts rather than raw audio queries. Within FIRE, early CLIR tasks were purely text-based, and only recently have shared tasks such as SqCLIR 2024 [8, 9] introduced real spoken queries for Indic languages. More recently, the study SqCLIRIL [10] further advanced this direction by exploring end-to-end spoken query retrieval approaches for Indian languages, providing valuable insights that complement the present SqCLIR 2025 shared task [1].

The SqCLIR shared task series, organised under the FIRE initiative, aims to foster research in speech-driven retrieval for Indic languages. The second edition (SqCLIR 2025) introduced several key enhancements over the first edition. We adopted the IndicMSMARCO [11] dataset as the retrieval collection, providing a large-scale, multilingual benchmark suitable for both monolingual and cross-lingual retrieval. Spoken queries were derived from TREC DL'19 [12] and DL'20 [13] queries and re-recorded in multiple Indic languages to simulate realistic voice-search conditions. The task expanded to five languages—Gujarati, Hindi, Bengali, Kannada, and English—making it one of the most comprehensive spoken IR evaluations in the Indic context.

The primary objectives of SqCLIR 2025 were to:

- Establish a benchmark platform for evaluating spoken query retrieval systems across multiple Indic languages.
- Encourage research in monolingual and cross-lingual spoken information retrieval.
- Provide standardised datasets and evaluation protocols to ensure reproducibility and fair comparison.
- Identify key challenges in integrating speech recognition, translation, and retrieval in low-resource multilingual settings.

To achieve these goals, two subtasks were designed to address distinct retrieval scenarios:

- **Task 1: Spoken Query Ad-Hoc Retrieval (Monolingual)** – Retrieve documents in the same language as the spoken query.
- **Task 2: Spoken Query Cross-Lingual Retrieval** – Retrieve documents written in a different target language from the spoken query.

Together, these subtasks aim to provide a unified and realistic benchmark for evaluating spoken query retrieval in multilingual Indic environments and to stimulate further research on speech-driven cross-lingual retrieval systems.

2. Task Definition

The SqCLIR 2025 shared task focuses on developing and evaluating systems that can effectively retrieve relevant textual documents given spoken queries in Indic languages. Participants are provided with a text-based query along with its corresponding spoken version in wav format. They are encouraged to utilize the provided spoken queries or generate additional spoken samples recorded under varied environmental conditions to test system robustness. The shared task is divided into two subtasks designed to evaluate both monolingual and cross-lingual retrieval capabilities.

[1] https://sites.google.com/view/sqclir-2025

2.1. Task 1: Spoken Query Ad-Hoc Retrieval (Monolingual)

In this subtask, participants are required to develop a spoken query retrieval system capable of handling monolingual queries. Both the spoken query and the target document collection belong to the same language, making the retrieval process comparatively straightforward. The objective is to accurately interpret spoken queries and retrieve the most relevant documents from the text corpus in the same language. For SqCLIR 2025, the monolingual task includes five Indic and English languages: Gujarati, Hindi, Bengali, Kannada, and English.

2.2. Task 2: Spoken Query Cross-Lingual Retrieval

The second subtask focuses on cross-lingual retrieval, where the spoken query and the document collection are in different languages. Participants are required to design retrieval systems that can interpret a spoken query in one language and return the most relevant documents written in another language. This task introduces additional challenges such as translation ambiguity, speech recognition errors, and cross-lingual semantic alignment. The task involves the same five languages—Gujarati, Hindi, Bengali, Kannada, and English—and allows for any combination of query–document language pairs. This flexible setup enables participants to explore a variety of cross-lingual retrieval strategies and evaluate system performance under multilingual and speech-driven conditions.

3. Dataset and Resources

The SqCLIR 2025 shared task employed large-scale multilingual datasets for text retrieval and spoken query evaluation across Indic languages. The primary text collections provided to the participants were the IndicMSMARCO [11] dataset for Gujarati, Hindi, Bengali, and Kannada, and the original MSMARCO [14] passage ranking dataset for English. IndicMSMARCO, an extension of MSMARCO, offers a multilingual benchmark consisting of query document pairs translated from English into multiple Indic languages. It comprises over 8.8 million passages, enabling consistent evaluation across both monolingual and cross-lingual retrieval tasks. The inclusion of both Indic and English collections ensured that retrieval experiments could be performed under realistic conditions, supporting a wide range of research scenarios within the SqCLIR 2025 task.

The spoken query dataset was derived from the TREC DL'19 and DL'20 query sets, consisting of a total of 97 queries. The translated and recorded spoken queries were sourced directly from the SqCLIRIL [10] study, which originally prepared these resources for spoken query cross-lingual retrieval research. For each language: Gujarati, Hindi, Bengali, Kannada, and English, one male and one female speaker's recordings were taken from SqCLIRIL to ensure balanced gender representation and natural variability in pronunciation and acoustic characteristics. The recordings were distributed in wav format with consistent sampling rates and controlled durations. To further test system robustness, participants were encouraged to record their own spoken versions of the same queries in the specified format and under varied acoustic conditions for selected languages, as outlined in the task guidelines.

All datasets and resources were made available through the FIRE 2025 SqCLIR portal. The release package included the text queries, spoken query files, and query relevance judgments (qrels) for evaluation. Together, these datasets and tools form a comprehensive benchmark for advancing research in spoken query and cross-lingual information retrieval for low-resource Indic languages.

4. Evaluation Setup

For both tasks, system performance was evaluated using standard information retrieval metrics. The primary evaluation metric was nDCG@10, which measures ranking quality based on graded relevance. Additional metrics included MAP, MRR, Recall@100, and Recall@1000, providing a comprehensive view of retrieval effectiveness across different evaluation depths. This metric suite ensured consistent

and comparable assessment of spoken and cross-lingual retrieval systems submitted to the SqCLIR 2025 shared task.

5. Results and Discussion

A total of six teams registered for the SqCLIR 2025 shared task; however, only one team submitted a valid run corresponding to the Hindi monolingual track. The submitted system evaluated retrieval performance using both traditional and neural baselines, including BM25 and IndicBERT, across four query conditions: text queries, spoken queries (for males and females), and participant-recorded spoken queries.

As presented in Table 1, the results obtained by the participant team [15] were the best retrieval effectiveness achieved for text queries, which represent the upper bound of system performance in the absence of speech-related errors. For BM25, text queries achieved an nDCG@10 of 0.2024 and a MRR of 0.4497, while IndicBERT achieved an nDCG@10 of 0.1638 and an MRR of 0.3618. In contrast, performance for spoken queries was substantially lower due to acoustic variability and ASR errors. Among spoken inputs, the female query recordings consistently outperformed male recordings across both retrieval models, possibly due to clearer articulation and less background noise in the recorded samples. The participant-recorded queries produced comparable results to the provided spoken queries, indicating a degree of robustness in the evaluation design.

Despite limited participation, the results provide a valuable reference point for future research in spoken information retrieval for Indic languages. The performance gap between text and spoken queries highlights the challenges of ASR accuracy, pronunciation diversity, and domain mismatch core issues that must be addressed to advance cross-lingual speech-based retrieval in low-resource settings.

Table 1
Participant team results on the Hindi SqCLIR Task using BM25 and IndicBERT models

Query Type	Model	nDCG@10	MRR	R@100	R@1000
Text query	BM25	0.2024	0.4497	0.1767	0.3358
	IndicBERT	0.1638	0.3618	0.1241	0.2608
Spoken query (female)	BM25	0.0951	0.2462	0.1050	0.1976
	IndicBERT	0.0715	0.1964	0.0674	0.1475
Spoken query (male)	BM25	0.0751	0.1872	0.0782	0.1834
	IndicBERT	0.0586	0.1323	0.0640	0.1348
Participant Recorded query (male)	BM25	0.0814	0.1992	0.0776	0.1584
	IndicBERT	0.0744	0.1786	0.0629	0.1371

6. Conclusion

The SqCLIR 2025 shared task advanced research in spoken query–based cross-lingual information retrieval for Indic languages, extending the first edition by adopting the large-scale IndicMSMARCO dataset and incorporating spoken queries across five languages—Gujarati, Hindi, Bengali, Kannada, and English. Despite limited participation, the task successfully established a benchmark framework for evaluating both monolingual and cross-lingual retrieval using spoken input. The findings reveal that text queries consistently outperform spoken ones, largely due to automatic speech recognition errors, pronunciation variability, and background noise. Nonetheless, SqCLIR 2025 demonstrates the feasibility of large-scale spoken query retrieval in low-resource Indic settings and provides a strong foundation for future work focusing on robust ASR integration, cross-lingual modeling, and end-to-end neural speech retrieval systems.

Acknowledgments

We sincerely thank the organizers of FIRE 2025 for providing the opportunity to host the SqCLIR track as part of the conference. We also express our heartfelt gratitude to the native speakers who contributed to the creation of the spoken query dataset—their support and dedication were instrumental in the successful development of this resource.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. doi:10.1561/1500000019.

[2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[3] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2020, pp. 39–48. doi:10.1145/3397271.3401075.

[4] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic bert sentence embedding, arXiv preprint arXiv:2007.01852 (2022). URL: https://arxiv.org/abs/2007.01852.

[5] D. W. Oard, A. R. Diekema, The state of the art in cross-language information retrieval, in: Proceedings of the American Society for Information Science Annual Meeting, volume 35, 1998, pp. 237–246.

[6] G. Grefenstette, Cross-Language Information Retrieval, Springer Science & Business Media, 2012. doi:10.1007/978-94-011-5206-1.

[7] J. S. Garofolo, E. M. Voorhees, V. Stanford, D. S. Pallett, W. M. Fisher, N. L. Jones, The trec spoken document retrieval track: A success story, in: Proceedings of the Content-Based Multimedia Information Access Conference (RIAO), 2000, pp. 1–20.

[8] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Overview of the fire 2024 sqclir track: Spoken query cross-lingual information retrieval for the indic languages, in: Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2024) Working Notes, volume 3810, CEUR-WS.org, 2024, pp. 1–8. URL: https://ceur-ws.org/Vol-3810/paperXXX.pdf, available at CEUR Workshop Proceedings.

[9] B. Dave, P. Majumder, D. Ganguly, E. Kanoulas, Findings of shared task on spoken query cross-lingual information retrieval for the indic languages at fire 2024, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '24, Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3734947.3735669.

[10] B. Dave, P. Majumder, Sqcliril: Spoken query cross-lingual information retrieval in indian languages, Pattern Recognition Letters (2025). URL: https://www.sciencedirect.com/science/article/pii/S0167865525003071. doi:https://doi.org/10.1016/j.patrec.2025.08.022.

[11] S. Haq, A. Sharma, O. Khattab, N. Chhaya, P. Bhattacharyya, IndicIRSuite: Multilingual dataset and neural information models for Indian languages, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume

2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 501–509. doi:10.18653/v1/2024.acl-short.46.

[12] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 deep learning track, CoRR abs/2003.07820 (2020). URL: https://arxiv.org/abs/2003.07820. arXiv:2003.07820.

[13] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, CoRR abs/2102.07662 (2021). URL: https://arxiv.org/abs/2102.07662. arXiv:2102.07662.

[14] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, CoRR abs/1611.09268 (2016). URL: http://arxiv.org/abs/1611.09268. arXiv:1611.09268.

[15] P. TT, T. KK, B. B, Spoken-query cross-lingual information retrieval for the indic languages (sqclir) using bm25 and indic-bert (2025).