International Journal of Convergent Research 2(2) July - December (2025)

IJCR
International Journal of Convergent Research

International Journal of Convergent Research

Journal homepage: International Journal of Convergent Research

Speech-Driven Information Retrieval Systems: A Review of AI and NLP Techniques

Hamza Aziz[1], Heena Yousuf [2]

[1] University of Pisa, Italy
[2] Department of Computer Science, Aligarh Muslim University, Aligarh, India

***Corresponding Author:** hamzaabdulaziz786@gmail.com

Citation: Aziz, H., & Yousuf, H. (2025). Speech-driven information retrieval systems: A review of AI and NLP techniques. *International Journal of Convergent Research, 2*(2). 8-17.

ARTICLE INFO

Received: 02nd December 2025
Accepted: 27th December 2025

ABSTRACT

This paper offers a critical review of the progress, issues, and future of speech-driven information retrieval (SDIR). Following ten studies as references, this paper focuses on the use of artificial intelligence (AI) and natural language processing (NLP) in support of speech-based interaction within information retrieval systems (IRSs). Innovative difficulties, including differences in the accents, noise, and the peculiarity of the words, and the contextual and multilingual approach, are also mentioned. This paper discusses how current trends in AI, such as transformers like GPT and BERT, restore in-depth features for the enhancement of the speech recognition rate, semantic content analysis, and operational queries in real time. Furthermore, future trends, including multilingual retrieval systems and real-time processing, are examined as significant advancements in improving the SDIR systems' accessibility and speed. Overcoming these challenges and building advances in AI, the study aims towards the development of future SDIR systems that offer optimal, easy, and versatile solutions for various uses.

Keywords: Speech-Driven Information Retrieval, Artificial Intelligence, Natural Language Processing, Speech Recognition, Transformer Models, Multilingual Retrieval.

INTRODUCTION

Recently, speech-driven information retrieval systems have been considered more frequently because of increasing requirements for effective, convenient, and natural ways of searching. These systems rely on voice as the primary method of user communications and convert words into searchable database queries, so the user does not have to touch a keypad or keyboard. Speech recognition with multimedia retrieval has provided new opportunities for user-centered systems that use voice triggers in a variety of contexts, such as voice-enabled assistants or speech-based search engines for the Internet or for specific contexts such as enterprise and healthcare information systems.

As part of AI and NLP technologies, there are significant opportunities for improvement in the SDIR systems. Advanced AI tools, specifically machine learning and deep learning, assist SDIR systems in analyzing speech data, thereby enabling them to incorporate a variety of accents, speech patterns, and contextual factors into their data analysis. By utilizing NLP techniques, these systems are able to capture the meaning behind spoken words and enhance advanced functions such as semantic search, intent recognition, and context-aware information search. Coupled with NLP, AI forms the basis of the success of SDIR systems; by meaning not only voice-to-text but understanding user inquiry to be specific, SDIR tools are ideal for real-time, dynamic, and conversational search.

FOUNDATION OF INFORMATION RETRIEVAL SYSTEMS

Traditional Information Retrieval Models

Information retrieval (IR) systems have been developed over the course of several decades; the systems initially matched words from the documents with the words entered by the users. The Boolean model, vector space model, and probabilistic model act as cornerstone architectures of most of the initial IR systems.

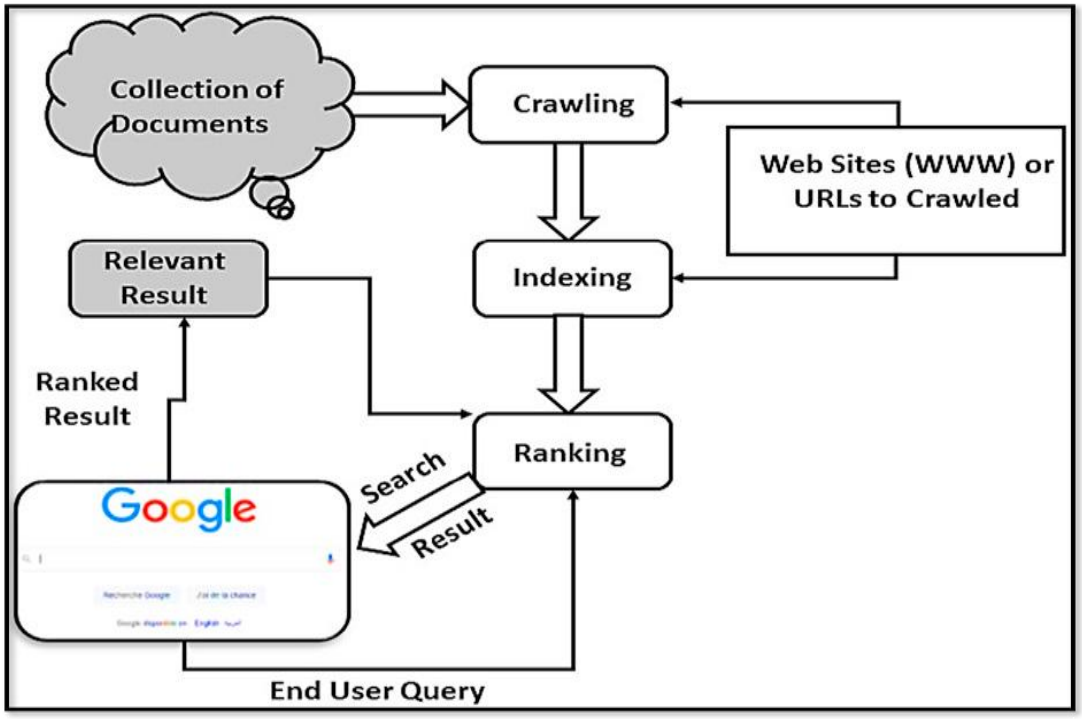
Boolean Model: A Boolean model works based on keywords, identifying documents that contain them or excluding them from the result set. While it is straightforward and very easy to apply, the Boolean model fails to capture the complexities involved in making queries and ranking output items based on relevancy. [12]

Vector Space Model: The vector space model extended the Boolean model by generating documents and queries in the form of vectors in a multidimensional space, with the possibility of a degree of similarity. This model incorporates the technique of calculation of term weights through the frequency of its appearance in a document and the number of documents in the corpus in which the given term appears. [13]

Probabilistic Model: The probabilistic model, unlike others, involves probabilistic models for estimating the relevance of the documents with reference to data acquired previously. They consider the likelihood of a document being relevant to a query, providing greater flexibility and reflection in the retrieval process. [14]

While seemingly cutting-edge, the traditional models are still predominantly textual and do not have native support for speech or other non-textual data.

Figure 1: *Information Retrieval Components [11]*



Integration of Speech in IR Systems

Incorporating speech into an information retrieval system introduces an additional layer of complexity, as the system converts speech into text for analysis. A spoken query has been used to convert spoken words to text through speech recognition technologies, which can be processed using conventional IR systems. Nonetheless, difficulties appear regarding some speech features like accent, speech rate, background noise, and the presence of homophones.

For the textual content of speech to be useful in an IR system, the spoken words need to be transcribed, but the AI and NLP algorithms must understand the query for it to work. This includes identifying user intent, dealing with vagueness, and dealing with context-relevant data. For instance, while the traditional text-based information retrieval systems work using a word-match approach, the speech-based information retrieval systems need to take into account the differences in slot-or-failure associated with spoken language.

The inclusion of speech also provides new search paradigms, such as voice and conversational agents, that are transforming the ways users interact with the existing IR systems. Synchronization of both speech recognition and advanced

NLP models can help the system enhance the user experience, improve query understanding, and provide the user with personalized user results as well as context-based results.

SPEECH RECOGNITION AND ITS ROLE IN SDIR SYSTEMS

Indeed, SDIR systems heavily rely on this technology since SDIR's basic purpose is to translate spoken language into a format that machines can read. There are techniques used together with the models that are used in the process of speech recognition, and all have a certain purpose of making the system better and better.

Techniques in Speech Recognition

Acoustic Models: Acoustic models are the establishment of the links between the phonetic units in any spoken language and the corresponding sounds. These models learn from large amounts of microphone recordings to establish the link between sounds and speech. Hidden Markov Models (HMMs) and Deep Neural Networks (DNNs) are common strategies that recognize small sections of the human voice, known as phonemes, and then string these phonemes into words.

Language Models: Language models make it easier with the recognition process because they give meaning to the decoded terms. They estimate the probabilities of the occurrence of particular sequences of words in a certain language; the recognition system is then enhanced as unlikely or non-existing word sequences are discarded. N-gram models, along with other more sophisticated language models such as recurrent neural networks (RNNs) and transformers, can capture syntax and grammar.

Feature Extraction: Suppose a fantastic speaker signals, there is much information that is not useful to the task of recognition. Preprocess methods include Mel-frequency cepstral coefficients (MFCCs), which change real audio data to features that can be better understood by the recognition system. These features reflect the main aspects of speech sounds, keeping aside any unnecessary noise or unwanted complexity.

End-to-End Speech Recognition: Most of the classical speech recognition systems are composed of independent parts for the acoustic modeling, the feature extraction, and the language modeling. However, current paradigms involve end-to-end models that integrate all components into a single deep learning model. DeepSpeech and WaveNet are examples of end-to-end models of speech recognition that take audio as input and spit out text as output.

By using these techniques, the SDIR systems can transcribe spoken queries into text form and search the information in the databases or over the web.

Challenges in Speech Recognition for IR

Despite the progress made concerning speech recognition, these deficits seem more pronounced when voice input is incorporated into information retrieval models.

Accents and Dialects: For example, one of the most compelling challenges is delivery, which includes issues with accents, dialects, and speech patterns across several speakers. Such variation can cause issues in speech transcription where the same combinations of symbols are read differently by potentially any person, even from a single country. Various images present in such applications may involve varying contrast and intensity, which the SDIR systems must be trained on different datasets to address such transformations.

Background Noise and Audio Quality: Disturbances are also present in the real-world paradigm in the form of noise that can be conversations, traffic noise, and other sounds occurring in the environment. This is arguably a familiar problem, particularly in mobile, or hands-free, scenarios in which users may find themselves speaking in loud places. One of the essential research areas is to build reliable systems that are capable of detecting speech in noisy environments. [16]

Ambiguity and Homophones: In spoken language, authors occasionally employ homophones, which share a similar sound but differ in spelling and/or meaning (e.g., 'sea' and 'see'). Thus, the following are examples of samples that cause difficulty in transcription, since more often than not, the correct interpretation depends on the context in which they are said. The role of context and power words is obvious, and only the most sophisticated NLP models can decode these homophones.

Real-Time Processing: As mentioned before, for SDIR systems to be fully interactive and easily usable, the systems must be able to perform speech recognition in real time. This requirement makes the need for efficient speech recognition models that can work within very tight constraints of time while maintaining accuracy. Delayed responses can also be a disadvantage in present-day systems since adaptability directly correlates with response time.

Long and Complex Queries: The format of spoken queries may be as simple as a few keywords typed in a search engine, but it may be longer, more complex, and less carefully structured. Clients tend to be more wordy; they use more informal language or employ filler words; they also may ask several questions at once. This could potentially complicate the process of identifying the speech's motivation in relation to the documents stored in an IR system. Regarding such complex queries, advanced NLP models such as the transformers and BERT-based architecture are applied that take into account semantic relevance and context.

These are the main challenges that must be tackled to achieve improvements in the highly effective SDIR systems.

Optimizing accent variation, background noise, or the possibility to process the information in real time can further enhance the general usability of SDIR systems and widen the sphere of their application.

NATURAL LANGUAGE PROCESSING TECHNIQUES FOR SDIR

Speech-to-Text Conversion and NLP

Speech recognition is the initial process in most of the Speech-Driven Information Retrieval (SDIR) systems that translate voice input into text for further processing. And this conversion is done with the help of automatic speech recognition (ASR) systems, which rely on acoustic and language models to recognize spoken words easily. However, the plain text format generated by transcription is not sufficient to support information retrieval since the content may be noisy and possibly disruptive for processing.

Here, natural language processing (NLP) also has a part and helps to filter the transcribed text. NLP tasks in this stage are as follows:

- Text normalization: To preprocess the text, clean it and make it standard by eliminating such words as “and” and “or” and making general corrections, such as changing informal and incorrect writing to standard English.
- Part-of-speech tagging: Adding tags with the grammatical functions of the system to provide a better understanding of all the words' positions in a sentence.
- Tokenization: Splitting the text into smaller units (e.g., words or phrases) for further analysis.
- Query formulation: Structuring the text into a formal query format suitable for the information retrieval system.

By integrating accurate speech recognition with practical NLP processes, SDIR systems can process the received user queries and convert them into useful real language interpretation for information retrieval input.

Named Entity Recognition (NER) in Speech Data

Named Entity Recognition (NER) is an essential NLP application to recognize and extract entities like names, place names, dates, and other proper terms from transcribed speech. In verbal queries, due to advanced processing of voice, the NER helps further progress of the SDIR system in providing more relevant information. [17]

For example, in a query like “Show me hotels near Central Park,” NER identifies "Central Park" as a location entity, allowing the system to narrow the search context accordingly.

Key benefits of NER in SDIR systems include:

- Improved query accuracy: By isolating and categorizing entities, NER ensures that the most relevant keywords are prioritized in the search process.
- Contextual disambiguation: Resolving ambiguities, such as distinguishing between homonyms (e.g., "Apple" as a company or fruit).
- Query expansion: Using the recognized entities to the inclusion of synonyms or related terms to get broader and more accurate results.

Other states, like BERT and RoBERTa-based transformers in deep learning, have made remarkable performance in NER regardless of the noisy or incomplete speech data. Despite challenges such as accentuation and speech errors, NER continues to be a fundamental method for extracting valuable information from speech inputs.

Sentiment Analysis and Semantic Understanding

Another type of information that SDIR systems focus on is the emotional tone and contextual intent in the user's request. Two NLP techniques in particular meet this requirement: sentiment analysis and, to a greater extent, semantic understanding.

Semantic Analysis: It examines the emotional tone (e.g., positive, negative, or neutral) expressed in a spoken inquiry. A query such as “What are the best-rated restaurants near me?” shows a positive sentiment in search of superior possibilities, whereas “Which hotels have the worst reviews?” indicates a negative feeling.

Applications of sentiment analysis in SDIR include:

- Tailoring search results based on the detected sentiment
- Enhancing user interaction by responding empathetically or contextually.

Specific difficulties in sentiment analysis for speech data include the identification of some specific nuances, like sarcasm, or when the speaker is expressing different emotions at the same time, or some changes of speech tone that affect the sentiment. [15]

Semantic Understanding: Semantic understanding dwells on the true meaning as well as the real reason behind spoken queries. In contrast to traditional Boolean search, which strictly depends on word-to-word matching, semantic methods also

search for the relevance between words and context to yield better results.

Techniques in semantic understanding include:

- Word embeddings: Representing words in vector spaces to capture their meanings and relationships (e.g., Word2Vec, GloVe).
- Transformer models: Advanced models such as BERT and GPT examine word sequences and contexts to improve query comprehension.
- Intent recognition: Identifying the specific type of query, such as informational (e.g., “What’s the weather today?”), navigational (e.g., “Find Starbucks near me”), or transactional (e.g., “Book a flight to New York”).

The above examples demonstrate that by using sentiment analysis and semantic understanding, SDIR systems can go beyond the core instructions from a vocal query to provide the context and user-based results. These techniques enable SDIR systems to not only listen to the word the user spoke but also analyze the message he actually wanted to convey, which is necessary in order to make further steps towards more natural and effective techniques of information retrieval.

LITERATURE REVIEW

Citation: S. Ibrihich, A.Oussous, O. Ibrihich, M. Esghir A, “Review on recent research in information retrieval”

Brief Summary- In the following paper, an initial overview of modeling and simulation for describing the fundamentals of information retrieval. Some of the issues and techniques relating to methods used in a system, challenges faced, models to be adopted, and components of an IR system are examined. Much of this paper is drawn from the paper where the writer discusses some of the common terms with reference to the information retrieval system.

Citation: Singhal. A, “Modern Information Retrieval”

Brief Summary- This article is an attempt to present an outline of the major progress achieved in the information retrieval area, as well as to indicate what the state of the art is at present.

Citation: de Campos, L.M.; Fernández-Luna, J.M.; Huete, J.F.; Ribadas-Pena, F.J.; Bolaños, N., “Information Retrieval and Machine Learning Methods for Academic Expert Finding”

Brief Summary- This paper focuses on academic expert finding by analyzing the relevance of using information retrieval (IR) and machine learning (ML) techniques, of which deep learning is one.

Citation: C. González-Ferreras and V. Cardenoso-Payo. "A system for speech-driven information retrieval." 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), pp. 674–679, Kyoto, Japan, Dec. 2007. DOI: 10.1109/ASRU.2007.4430184

Brief Summary- This paper presents an approach to information seeking from a document collection by voice in Spanish, that is, spoken queries. The system links a speech recognizer with an information retrieval engine. It uses accommodations in Word and language modeling to solve the out-of-vocabulary (OOV) word problem and lessen word error rates (WER). Also, under non-English-speaking environments, a pronunciation lexicon expansion was introduced as a way of raising the performance level. We found out that using the CLEF’01 test set, the retrieval precision has improved by 6.34% relative to before, the collected sets contain 24.71% fewer OOV words, and the sets have a WER of 10.87% less than before, proving the effectiveness of the spoken query processing in augmenting the existing system.

Citation: F. Crestani, "Spoken query processing for interactive information retrieval," Data & Knowledge Engineering, vol. 41, no. 1, pp. 105–124, Feb. 2002. DOI: 10.1016/S0169-023X(02)00024-1

Brief Summary- This paper aims at discussing the effectiveness of the spoken query processing on the interactivity of Information Retrieval (IR) systems. Taking advantage of the recent developments in automatic speech recognition, the paper assesses the impact of word recognition errors in spoken queries on classical IR techniques. Experiments show that these re-ranking techniques are effective and still maintain their high accuracy rates even as the error rate of the documents for long queries. However, for short queries, the quality of the spoken query processing can be boosted notably with the help of both standard and pseudo-relevance feedback. This paper strengthens the practicality of using speech as a means of interaction with IR systems while recognizing issues of query length and misrecognition.

Citation: A. Fujii, K. Itou, and T. Ishikawa, "Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition," in Information Retrieval Techniques for Speech Applications (LNCS 2273), pp. 94–104, Springer, 2002

Brief Summary- This research work introduces a new idea about how to use speech to retrieve text by combining a speech recognition system with retrieval technology. The method improves both recognition and spoken query retrieval effectiveness in target collections by incorporating statistical language models tailored for such collections. A verification of the approach is done with experiments using existing test collections and dictated queries that show the effectiveness of the new approach for

improving the precision and recall of speech-based retrieval systems.

Citation: M. V. Mahajan and X. D. Huang, "Information Retrieval and Speech Recognition Based on Language Models," Microsoft Technology Licensing LLC, Patent

Brief Summary- In this research, information retrieval and speech recognition are combined under language models. The system described uses a two-datastore model, with one smaller data store to construct queries for the larger data store and changing language models in real time. The language models derived from these data stores are used for speech recognition, as well as for document look-up. The system computes document perplexities based on these models to rank them for relevancy and achieve documents that are above a certain relevancy level. This method improves the retrieval accuracy while also improving the sophistication of speech recognition.

Citation: C. González Ferreras and V. Cardeñoso-Payo, "Dynamic Adaptation of Language Models in Speech-Driven Information Retrieval," TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. DOI: 10.1007/978-3-540-74628-7_29

Brief Summary- This paper assesses a system that takes voice commands and uses these to search for information in a text document base. In order to handle spoken queries, it uses a continuous speech recognizer with a large vocabulary for converting the spoken word into text, and apart from it, there is an information retrieval system that searches for the particular document. Two-pass systems can be efficiently enhanced using a dynamic approach to language models. The implemented system is for the Spanish language, and it was evaluated by carrying out experiments using the CLEF IR test suite with recorded spoken queries of 10 users. Achievements show that the model, with 60,000 words, improved the index recall rate, including retrieval precision, by 5.74% from the baseline.

Citation: K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond".

Brief Summary- The following paper presents a brief overview of the information retrieval models used in primary and subsequent steps of the processing line. It includes current models based on terms, semantic search, and even neural techniques. Furthermore, it overviews specific educational activities related to the values reflected by these models. It reveals features of the learning process that can be helpful for other researchers and practitioners in the sphere of information search.

Citation: Manal Sheikh Oghli, Muhammad Mazen Almustafa, "Comparison of basic Information Retrieval Models"

Brief Summary- In light of these challenges, this paper emphasizes the need to come up with efficient information retrieval models to cater to this demand. The paper also discusses simple models of information retrieval, with preference given to the vector space model. Although the model is commonly applied, there appear to be some difficulties, for example, the variety of approaches to defining term weights and the assumption of termhood. Overcoming such loopholes might improve its performance in the execution of the retrieval missions.

ANALYSIS OF PAPERS

Altogether, ten papers were used as references for this paper. The comparison table captures the various paper titles, their paper type (survey, conceptual, or experiment), as well as the techniques used for their writing, for instance, literature review or experimental design.

A summary of results and the limitations of the papers is mentioned, along with the references.

Table 1: *Comparison table*

Paper	*Features*				
	Paper type	*Technique used*	*Summary of results*	*Summary of limitation*	*Reference*
A Review of Recent Research in Information Retrieval	Survey	Literature Review	Complete information about basic terminology in information retrieval systems.	More new technologies are yet to come in the field of Information Retrieval.	[1]
Modern Information Retrieval	Conceptual	Experiment	Modern IR systems being developed continuously are leading to efficient web search engines.	IR systems are still identifying the different problems faced by users.	[2]
Information Retrieval and Machine Learning Methods for Academic Expert Finding	Survey	Literature Review	IR systems based on neural networks are useful in training ML Models.	In the development of a user-based recommendation system	[3]

A system for speech-driven information retrieval	Experiment	Speech Recognition with Adapted Vocabulary and Language Model; Pronunciation Lexicon Expansion	Improved retrieval precision (6.34% relative gain), reduced out-of-vocabulary (OOV) word rate (24.71% relative reduction), and lower word error rate (WER) (10.87% relative reduction).	Limited to the Spanish language; system performance may vary with different languages or accents; requires extensive adaptation for multilingual support.	[4]
Spoken query processing for interactive information retrieval	Experiment	Classical Information Retrieval Techniques, Relevance Feedback, and Pseudo-Relevance Feedback	Demonstrated robustness of IR systems to high levels of word recognition errors for long spoken queries. Relevance feedback methods improved effectiveness for short queries.	System effectiveness may degrade with extremely high error rates in short queries without feedback mechanisms; reliance on classical IR techniques limits adaptability.	[5]
Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition	Experiment	Statistical Language Model Adaptation for Speech Recognition and Integration with Information Retrieval Methods	Enhanced recognition and retrieval accuracy by adapting statistical language models to the target IR collection. Effectiveness demonstrated through experiments with test collections.	The approach relies heavily on the quality of target collections; it may face challenges with diverse or dynamic datasets and spoken queries not closely aligned to the collection.	[6]
Information retrieval and speech recognition based on language models	Patent / Methodology	Language Model Adaptation, Perplexity-Based Relevance Assessment	Proposes a dual-dataset approach to adapt language models for speech recognition and information retrieval, improving relevance and retrieval precision.	The approach heavily relies on perplexity thresholds, which may not generalize well to highly diverse or dynamic datasets.	[7]
Dynamic Adaptation of Language Models in Speech-Driven Information Retrieval	Experiment	Dynamic Language Model Adaptation, Two-Pass Retrieval Approach	Demonstrates improved retrieval precision (5.74% gain) using dynamic language model adaptation for Spanish spoken queries in a textual document collection.	Limited to the Spanish language and evaluated on a specific dataset; performance may vary with other languages and larger vocabulary sizes.	[8]
Information Retrieval: Recent Advances and Beyond.	Survey	Literature Review	Overview of the semantic retrieval models in the context of information retrieval.	It highlights the challenges and difficulties in the field.	[9]
Comparison of Basic Information Retrieval Models	Conceptual	Literature Review	The VSM is considered the most flexible and clear to date.	To increase the effectiveness of the terms weighing process by defining descriptors of terms in documents, to overcome the weaknesses in VSM.	[10]

CHALLENGES AND LIMITATIONS OF SPEECH-DRIVEN IR

Altogether, ten papers were used as references for this paper. The comparison table captures the various paper titles, their paper type (survey, conceptual, or experiment), as well as the techniques used for their writing, for instance, literature review or experimental.

Accents, Noises, and Ambiguities in Speech Data

It is therefore clear that one of the biggest problems for SDIR systems is the variation found in spoken data. Regional dialects and regional accents both affect the resultant speech recognition in a very great way. In fact, there are a lot of differences

in people's pronunciation. If the developed system were trained on a small set of samples with a specific accent, then a rough accent might lead to a wrong transcription. It becomes worse in the areas where people use multiple languages, and depending on the dialectical difference, the phrases used may be very different. In addition, there is a constantly varying noise level that additionally makes it difficult to get good voice input. Speaking environments ranging from public spaces, working places, or even apartments or houses with background noise interfere with speech signals and hence cause many misunderstandings during transcription.

Beyond accents and noise, another challenge lies in the use of imprecise language. People often speak in written queries that contain fragments of a reconstructed conversation, which also makes it difficult to construct the sentence, homophones, or colloquialisms. For example, some homonyms, like write and right, or some ambiguous phrases, might create problems for the system. These limitations lower the effectiveness of SDIR systems since they allow outputting unjustified or suboptimal search results. To address these problems, it is necessary to use better acoustic models, noise suppression, and data that would involve variability on the linguistic and acoustic levels.

Language and Contextual Understanding

Understandably, the identification of the discrete components of speech in a given context is crucial for the function of the SDIR systems but remains complex. As in many other areas, users tend to primarily switch between languages in a single query, which is called code-switching. This practice increases the degree of challenge in both the actual speech recognition and the overall natural language processing that comes after this. Subsequently, when the system is not created with capabilities to handle multiple languages, queries involving multilingual inputs can be challenging for transcription or interpretation, consequently reducing the applicability of the SDIR systems in multicultural environments.

Contextuality is also one of the major challenges of using models in design. Conversational queries are spoken, hence they provide vague information about what users are searching for. For example, a question like "What's the temperature?" needs context data, including the user's location, to offer accurate results. Due to this, the kind of contextual metadata, like user history or geolocation, is very hard for SDIR systems to determine the exact intent of a user. Also, speech contains features that could be challenging for a system to understand because they bear certain meanings, for instance, idiom, irony or tone. For instance, the self-assertion "That's just great" might be interpreted in two ways: as either the speaker really meaning it and being genuinely happy and satisfied, or he was annoyed and was being sarcastic.

Overcoming these challenges is only possible by improving the technological approach to multilingual processing, better context assessment, and semantic analysis. As with any technological advancement, the social aspect of speech includes variability and a thorough complication of human speech, meaning that advances continue to be made through adding natural language processing and machine learning to speech recognition. By realizing these limitations, the SDIR systems may be adopted with greater solidity and made available for a more extensive range of users while operating in different contexts.

EMERGING TRENDS AND FUTURE DIRECTIONS

Role of Transformer Models and Pre-trained Language Models

The modern architectures of transformers like GPT, BERT, and Whisper are making great breakthroughs for SDIR to develop better contextual and semantic search. Relevance feedback and features of query interpretation and relevance ranking are remarkably improved due to the large-scale data in pre-trained models. Both claimed that the integration of their approaches with ours will enhance the reliability and flexibility of SDIR systems.

Multilingual Speech-Driven Information Retrieval

With the increasing need for cross-lingual indexing, the use of SDIR, which is available in multiple languages, is now required. Cross-lingual embeddings mBERT and XLM-R allow obtaining accurate processing of various languages and code-switching. Future systems should have the ability to support linguistic differences in different regions throughout the world and language translation between them.

Real-time Speech Processing for IR

Real-time SDIR concerns low-latency time aimed at providing nearly immediate responses to requests based on streaming systems and edge computing. These advancements make the speech-to-query practical because, as users speak, the system enhances natural interactions and the user experience. Real-time features will further enhance speed and contextuality in SDIR systems as soon as this functionality is developed.

CONCLUSION

Thus, this paper aimed at identifying the importance of AI and NLP to the development of speech-driven information retrieval (SDIR) systems. It emphasized basic features of old-school IR systems, how speech-based inputs were incorporated, as well as how stronger recognition and NLP strategies were employed for better query interpretation. Speech ambiguity, noise, real-time processing, and multilingual processing were discussed along with research opportunities, including transformer-based models. These performances strongly indicate the role of SDIR in stimulating advanced discovery of precise, pertinent, and easily accessible information.

Solutions for overcoming current limitations of SDIR, therefore, include the use of deep learning, multilingual models, and real-time processing enhancements to further advance the field into the future. Developments in the last few years have indicated that speech recognition, contextual interpretation, and individual user requests have all improved in accuracy and relevance, and trends indicate adoption in a broader range of applications than just personal assistants. As further research is done on SDIR systems, it will become more liberal, precise, and easier to understand, making the gap between natural language and the system small.

Ultimately, a truly smart city does not automate governance but one that humanizes it—a city wise enough to ask who is being left behind, and courageous enough to redesign itself in response. This paper offers not only a framework for understanding these challenges but a call to action: to build cities that are not just intelligent but inclusive, not just digital but democratic, and above all, not just smart but just.

ETHICAL DECLARATION

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this paper.
Financing: This research received no external funding.
Peer review: Double anonymous peer review.

REFERENCES

[1] Ibrihich, S., Oussous, A., Ouafaa, I., & Esghir, M. (2022). “A Review on recent research in information retrieval.” Procedia Computer Science, 201, 777–782. https://doi.org/10.1016/j.procs.2022.03.106

[2] Singhal, A. & Google, Inc. (n.d.). “Modern Information Retrieval: A Brief Overview.” http://160592857366.free.fr/joe/ebooks/ShareData/Modern%20Information%20Retrieval%20-%20A%20Brief%20Overview.pdf

[3] L. M. de Campos, J. M. Fernández-Luna, J. F. Huete, F. J. Ribadas-Pena, and N. Bolaños, "Information retrieval and machine learning methods for academic expert finding," *Algorithms*, vol. 17, no. 2, p. 51, 2024. [Online]. Available: https://doi.org/10.3390/a17020051C.

[4] Gonzalez-Ferreras and V. Cardenoso-Payo, "A system for speech-driven information retrieval," 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), Kyoto, Japan, Dec. 2007, pp. 674–679, doi: 10.1109/ASRU.2007.4430184.

[5] F. Crestani, "Spoken query processing for interactive information retrieval," Data & Knowledge Engineering, vol. 41, no. 1, pp. 105–124, Feb. 2002, doi: 10.1016/S0169-023X(02)00024-1.

[6] Fujii, K. Itou, and T. Ishikawa, "Speech-Driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition," in Information Retrieval Techniques for Speech Applications (LNCS 2273), pp. 94–104, Springer, 2002.

[7] M. V. Mahajan and X. D. Huang, "Information Retrieval and Speech Recognition Based on Language Models," Microsoft Technology Licensing LLC, U.S. Patent 7,901,710, Feb. 2011.

[8] C. González Ferreras and V. Cardeñoso-Payo, “Dynamic Adaptation of Language Models in Speech-Driven Information Retrieval,” in Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, Sept. 3-7, 2007, pp. 241-248. DOI: 10.1007/978-3-540-74628-7_29.

[9] K. A. Hambarde and H. Proença, "Information Retrieval: Recent Advances and Beyond," IEEE Access, vol. PP, no. 99, pp. 1-1, Jan. 2023. DOI: 10.1109/ACCESS.2023.3295776. License: CC BY 4.0.

[10] Oghli, M. S., & Almustafa, M. M. (2021). “Comparison of Basic Information Retrieval Models”. International Journal of Engineering Research and Technology, 10(9). https://www.ijert.org/research/comparison-of-basic-information-retrieval-models-IJERTV10IS090092.pdf

[11] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, “A Review on recent research in information retrieval,” Procedia Computer Science, vol. 201, pp. 777782, 2022.