International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 75-81 www.ijaem.net ISSN: 2395-5252

Survey on Speech Recognition and Retrieval-Augmented Generation

Edwin Alex Shaji, Jerishab M, Leya Thomas, M Viraj Prabhu, Asst.Prof Chinchu M Pillai

Department of Computer Science College of Engineering Chengannur

Date of Submission: 01-12-2024 Date of Acceptance: 10-12-2024

ABSTRACT—Automatic speech recognition (ASR) and retrieval-augmented generation (RAG) systems have seen remarkable progress in handling multilingualism, noise robustness, real-time transcription, and knowledge-intensive tasks. The survey reviews 12 key papers that contribute to advancements in ASR and RAG, covering approaches like end-to-end multilingual models, noise-reduction techniques, and real-time speech processing. It also examines RAG systems that enhance generative models by integrating retrieval mechanisms for improved accuracy in tasks like question answering and summarization. By categorizing the papers into themes, this survey highlights key methodologies, compares their performance, and identifies future directions for improving ASR and RAG technologies in handling real-world challenges.

I. INTRODUCTION

Automatic speech recognition (ASR) and retrieval- augmented generation (RAG) have become pivotal technologies in natural language processing (NLP), enabling applications like voice assistants, real-time transcription, and open-domain question answering. These technologies have evolved significantly with the advent of deep learning, allowing models to handle a variety of complex tasks, from recognizing speech in noisy environments to retrieving accurate information for knowledge-intensive queries.

Despite these advancements, challenges remain, particularly in multilingual ASR, noise-robustness, and real-time speech processing. Moreover, the growing importance of knowledge retrieval for enhancing language models presents new opportunities to improve task-specific performance through RAG. The survey examines 12 key papers that explore advancements in ASR and RAG technologies, with a focus on advancing ASR systems and integrating retrieval mechanisms into language generation.

The papers selected for the survey represent significant contributions to the field of ASR and RAG, addressing a wide range of problems including multilingual support, noise reduction, and real-time transcription. A comprehensive review and analysis of these works provides insights into the current state of the art and identify potential directions for future research.

The survey categorizes the papers into the following themes:

- **Multilingual Speech Recognition**: Developing ASR models capable of handling multiple lan- guages without performance degradation.
- **Noise-Robust ASR**: Enhancing ASR systems to operate effectively in noisy environments.
- **Real-Time Transcription and Speech Processing**: Advancing techniques for accurate, real-time speech captioning and transcription.
- **Large-Scale Weak Supervision in ASR**: Utilizing weak supervision and large datasets to boost ASR accuracy.
- **Retrieval-Augmented Generation (RAG)**: Incorporating retrieval mechanisms to improve large language models (LLMs) in tasks like question answering.
- **General-Purpose Speech Summarization**: Summarizing spoken language directly using large language models.

The survey explores the methodologies, results, and challenges discussed in the selected works, offering insights into the evolution of ASR and RAG systems and identifying potential future developments needed to address ongoing challenges in the field.

A. Multilingual Speech Recognition

Advancements in multilingual ASR focus on handling multiple languages in a single system without sacrificing performance or accuracy. Key approaches involve joint training of multiple languages in one model and dynamic language switching for real-time applications. The models reviewed demonstrate improvements in recognition accuracy and support fluid transitions between languages, addressing the growing demand for multilingual support in speech recognition systems. A notable challenge remains in ensuring minimal confusion between similar languages, which is addressed by integrating language identifiers and leveraging language- specific features.

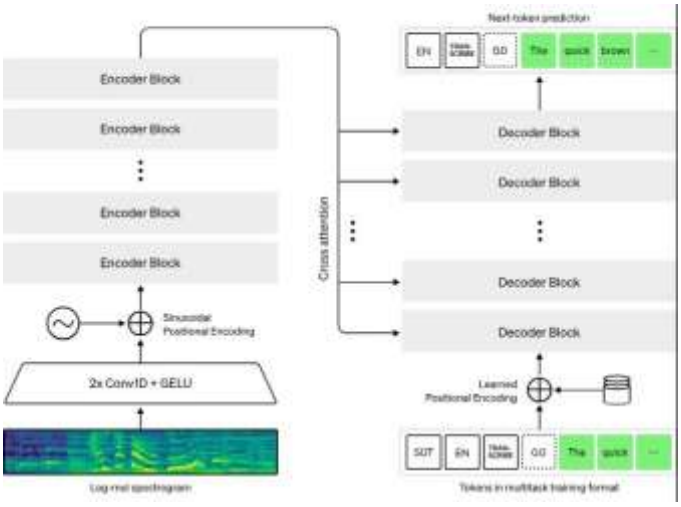

Fig.1:sequence to sequence learning in wisper

B. Noise-Robust ASR

Noise-robust ASR systems have become critical for real-world applications, where background noise can drastically affect performance. Techniques such as feature-domain processing (e.g., spectral subtraction)and model-domain approaches (e.g., neural networks) aim to reduce the impact of noise on ASR accuracy. Hybrid methods combining feature enhancement with advanced machine learning models have shown the most promise in mitigating the effects of noise. The trade- off between computational complexity and performance remains a central issue, especially in environments with non-stationary noise.

C. Real-Time Transcription and Processing

Real-time transcription systems are increasingly used in applications like live captioning and lecture transcription. Two main approaches have emerged: immediate feedback systems, which prioritize low-latency output, and post-processing systems, which optimize for accuracy after the event. Real-time systems face the challenge of balancing accuracy and speed, particularly when dealing with spontaneous speech or technical jargon. Innovations in audio splitting and the use of voice activity detection (VAD) have reduced latency while maintaining transcription accuracy, making real- time ASR more viable for practical deployment.

D. Large-Scale Weak Supervision in ASR

Weak supervision, particularly in the context of large datasets, has enabled substantial improvements in ASR systems. By training on hundreds of thousands of hours of weakly labeled data, ASR models can achieve near- human accuracy even in noisy or low-resource settings. In addition, quality improvements in human transcription, using confidence estimation and error correction models, contribute to creating better training datasets. These methods allow for scaling ASR systems to cover broader range of languages and dialects, making speech recognition accessible to more users worldwide.

E. Retrieval-Augmented Generation (RAG)

RAG models combine retrieval mechanisms with generative language models to improve performance on knowledge-intensive tasks. The retrieval process enhances the generation by providing relevant context or passages, enabling the model to generate more accurate and informed responses. RAG systems have been successfully applied to tasks like open-domain question answering and domain-specific knowledge retrieval, demonstrating superior performance compared to traditional methods. The use of dense embeddings and knowledge graphs further refines the retrieval process, allowing for more efficient and accurate information retrieval.

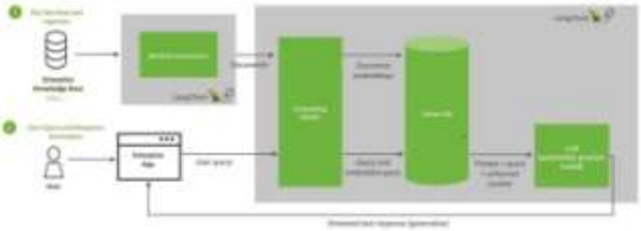

Fig.2: RAG pipeline

F. General-Purpose Speech Summarization

General-purpose speech summarization systems, which bypass traditional ASR steps by processing audio inputs directly, represent a significant advancement in the field. These systems use large language models (LLMs) combined with audio encoders to generate concise summaries of spoken content. By avoiding the errors introduced by ASR, such systems deliver more reliable and context-aware summaries across a variety of domains. This approach shows great promise for applications requiring real-time summarization and complex audio input processing.

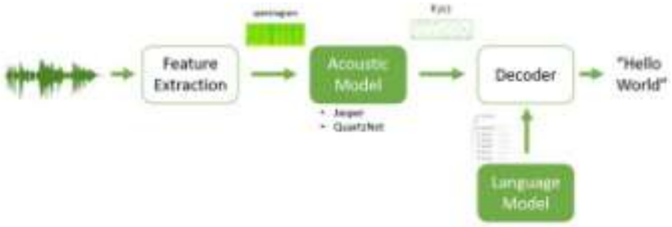

Fig.3: general purpose summarizarion asr model

II. LITERATURE REVIEW

Karpukhin et al. [1] propose an end-to-end sequence- to-sequence automatic speech recognition (ASR) model that handles multiple Indian languages with minimal script overlap. The model is trained on nine languages andoutperformslanguage-specificmodelsby21%. The inclusion of a language identifier helps to reduce confusion between similar languages, further improving accuracy.

Lietal.[2]review the progress in noise-robust ASR over the last 30 years, categorizing techniques into feature-domain and model-domain approaches. The paper emphasizes the need for hybrid methods combining both approaches and highlights the trade-offs between complexity and performance, particularly in handling uncertainty in noisy environments.

Zheng et al. [3] evaluate noise-reduction algorithms used for speech recognition, comparing spectral subtraction, MMSE, and model-based techniques. The authors conclude that model-based methods, particularly those using deep learning, achieve the best performance but are more computationally expensive than traditional methods.

Ranchal et al. [4] investigate real-time ASR for classroom settings, comparing real-time captioning (RTC) with post-lecture transcription (PLT). Their findings show that PLT offers better word recognition accuracy, especially in STEM courses, although RTC is preferred for immediate feedback.

Radford et al. [5] introduce Whisper, a speech recognition system trained on 680,000 hours of audio using weak supervision. The model approaches human- level transcription accuracy in low-resource languages and noisy environments, demonstrating the potential of large-scale weakly supervised learning for improving ASR systems.

Arriaga etal.[6] present a comparative evaluation of audio-splitting algorithms for real-time ASR. They explore voice activity detection (VAD), fixed-interval splitting, and a novel feedback algorithm. Their results show that VAD offers the highest transcription accuracy, while the feedback algorithm reduces latency without sacrificing significant accuracy.

Kang and Roy [7] propose a system that

integrates large language models (LLMs) with an audio encoder to perform speech summarization without relying on traditional ASR. Their model prompts LLMs directly with audio input, reducing the cascading errors typical of ASR-based systems and improving summarization performance across multiple domains.

Gonzalez-Dominguez etal.[8] describe a real-time ASR system that supports multiple languages through parallel speech recognition and language identification. The system allows dynamic switching between languages without significantly impacting recognition accuracy, making it ideal for multilingual users in real-time applications.

Karpukhin et al. [9] introduce Dense Passage Retrieval (DPR), a method that uses dense embeddings for passage retrieval in open-domain question answering tasks. DPR outperforms traditional retrieval methodslike BM25 and sets a new state-of-the-art in retrieval performance across multiple QA datasets.

Barronetal.[10] present SMART-SLIC, a framework that integrates retrieval-augmented generation with domain-specific vector stores and knowledge graphs. The framework excels in specialized tasks like malware analysis, using tensor factorization to improve the accuracy and relevance of the retrieved information.

Lewis et al. [11] explore the combination of retrieval- augmented generation (RAG) with dense passage retrievers and transformers. Their model achieves state- of-the-art performance on several knowledge-intensive tasks, including question answering and fact verification, highlighting the power of retrieval-based systems for improving language model outputs.

Gao et al. [12] address transcription errors in crowd sourced environments by introducing a confidence estimation model (CEM) and an error correction model (ECM).Their approach reduces transcription word error rates (TWER) by 50%, showing potential for enhancing ASR training datasets and improving overall transcription quality.

III. METHODOLOGYCOMPARISON

A. Model Architectures

The works surveyed employ various model architectures, from traditional pipeline models to modern end-to-end approaches. Pipeline models, though modular, suffer from error propagation, while end-to-end models combine all tasks into a single deep learning framework, enhancing performance, particularly in multilingual and real-time ASR. However, end-to-end models demand large datasets and significant computational resources.

In retrieval-augmented generation (RAG), integrating retrieval mechanisms into generative models introduces complexity, with dense passage retrieval (DPR) out performing traditional sparse retrieval. RAG models typically use a dual-encoder setup where the retriever and generator collaborate to improve response quality.

B. Training Techniques and Data Sources

Training approaches vary widely, with large-scale weak supervision enabling high-performance ASR, especially when large labeled datasets are unavailable. Multilingual ASR benefits from joint training across languages, with data augmentation improving model robustness against noise and unseen languages.

For RAG models, pre-trained language models like BART and GPT are fine-tuned for domain-specific tasks. Additional techniques such as tensor factorization and knowledge graph integration further enhance retrieval in domain-specific applications.

C. Evaluation Metrics

ASR systems are primarily evaluated using word error rate (WER), with language identification accuracy being crucial in multilingual systems. Noise-robust ASR uses metrics like signal-to-noise ratio (SNR) and spectral distance to gauge performance. In RAG models, retrieval accuracy and task-specific metrics (e.g., ROUGE, BLEU, METEOR) are key, alongside real-time factors (RTF) for computational efficiency.

D. Scalability and Practicality

End-to-end ASR models, while effective, require substantial resources and may be impractical in low- resource settings. Weak supervision allows scaling by leveraging large, weakly labeled datasets. Real- time transcription systems need careful optimization for latency and efficiency, balancing speed and accuracy.

RAG models, with dense retrieval mechanisms, scale well for large data sets but introduce computation a lover- head. Managing this overhead while maintaining real- time performance is crucial for large-scale applications.

E. Limitations and Trade-offs

End-to-end ASR models, though powerful, face difficulties with out-of-domain data and noise. Multilingual models may degrade when switching between similar languages. Noise-robust ASR systems also struggle to balance complexity with real-time performance.

In RAG systems, retrieval quality is critical, with errors leading to hallucinations or in correct outputs. Dense retrieval methods, while effective, require substantial memory and processing power, limiting their use in resource-constrained environments.

IV. CHALLENGES AND FUTURE DIRECTIONS

A. Challenges

- **Multilingual ASR Performance** Ensuring consistent performance across languages, especially for low-resource ones, is a challenge. Joint training improves overall accuracy but struggles with language switching and code-switching, often leading to errors. High-quality datasets for many languages are still limited, hindering training.
- **Noise-Robustness in Real-World Conditions** Handling complex, unpredictable noise remains difficult for ASR systems, particularly in mobile and real- time applications. Techniques like deep neural net- works offer better noise reduction but increase computational complexity, making them less practical for low-power devices.
- **Latency and Accuracy in Real-Time Transcription** Balancing low latency with high transcription accuracy is a key issue. Techniques like VAD help reduce latency but often sacrifice accuracy in spontaneous speech. Domain-specific fine-tuning for specialized jargon remains resource-intensive.

Focus Area	**Model Architecture**	**Data & Training Techniques**	**Key Metrics**	**Limitations**
Multilingual ASR	End-to-End Models with Language Identifiers	Joint training across an guages; data augmentation	WER, Language ID Ac- curacy	Struggles with similar languages and code- switching; requires extensive multilingual data
	Parallel Processing for Language ID and Recognition	Dynamiclanguage switching in real-time	WER, Language Switch Efficiency	Performance drop with rapid language switching in similar language sets
Noise-Robust ASR	Hybrid Models (Feature- Domain + Model- Domain Approaches)	Data augmentation with noise features	SNR, Spectral Distance, WER	High computational de-mands for complex noise environments
	Deep Learning-Based Noise Reduction	Supervised learning on clean and noisy data	Noise Reduction Accuracy, WER	Computationally intensive, limited real-time scalability
Real-Time Transcription	Immediate Feedback and Post-Processing Models	Audio splitting; Voice Activity Detection (VAD)	RTF, Latency, Word Recognition	Immediate feedback models often sacrifice accuracy for latency
	VAD with Adaptive Models for Real-Time Feedback	VAD and feedback mechanisms to reduce latency	Latency, Transcription Accuracy	VAD may decrease ac- curacy, especially with spontaneous or technical jargon
Weak Supervision in ASR	Large-Scale End-to-End Models	Trained on large, weakly labeled datasets	WER for low-resource languages	Requires extensive labeled data; lower performance on domain- specific speech
	Confidence Estimation and Error Correction Models	Crowd sourced data; error correction post- processing	TWER, Confidence Estimation	Limited impact in noisy conditions; latency impacted by error correction processes

Retrieval-Augmented Generation (RAG)	Dual-Encoder Models (Retriever + Generator)	Dense Passage Retrieval, Knowledge Graph Integration	Retrieval Accuracy, BLEU, ROUGE	High memory and processing overhead, prone to hallucination errors
	Domain-Specific RAG with Tensor Factorization	Domain-specific vector stores; tensor factorization	Task-Specific Accuracy	Complex to scale, requires significant computational resources for real-time performance
Speech Summarization	LLMs with Audio En- coders for Direct Summarization	Pre-trained LLMs fine- tuned on audio input	Summary Accuracy, ROUGE Score	Limited generalization across varied audio types; sensitive to ASR accuracy for input

TABLE I: Methodology Comparison by Focus Area

- **Scalability of RAG Models** While RAG models improve knowledge-intensive tasks, their scalability is limited by the computational over head of retrieval mechanisms. Large datasets and real-time performance require optimized retrieval systems to avoid errors like hallucination.
- **Hallucination in Generative Models** Generative models often produce factually incorrect outputs. This is a critical issue in high-stakes applications like question answering or legal and medical retrieval, where ensuring accuracy and attribution of knowledge is essential.

B. Future Directions

- **Improving Multilingual ASR for Low-Resource Languages** Cross-lingual transfer learning and data augmentation can enhance training for low-resource languages. Collaborative efforts to build larger multrilingual datasets are necessary to improve performance across diverse languages.
- **Advances in Adaptive Noise-Robust ASR** Future ASR systems should incorporate adaptive noise- cancellation methods and self-supervised learning to handle changing noise environments. Hardware- optimized algorithms will be crucial for mobile and edge devices.
- **Low-Latency Real-Time Transcription** Research should focus on hybrid models that can switch between high-accuracy and low-latency modes. Efficient fine-tuning techniques for domain-specific applications, like technical or medical fields, will also be essential.
- **Optimizing RAG Models for Large-Scale Applications** Future work should focus on refining dense retrieval techniques and integrating knowledge graphs more effectively to enhance accuracy and scalability, particularly for real-time question answering and interactive systems.
- **Fact Verification in Generative Models** Reducing hallucinations and improving fact verification will be key. Models should include stronger fact- checking mechanisms and provenance tracking to ensure the integrity and transparency of generated content, especially in high-stakes domains.

V. CONCLUSION

Automatic speech recognition (ASR) and retrieval- augmented generation (RAG) systems have made significant advancements in recent years, particularly in multilingual support, noise robustness, real-time processing, and handling knowledge-intensive tasks. The integration of large-scale weak supervision, noise reduction techniques, and dense retrieval mechanisms has pushed the boundaries of what ASR and RAG systems can achieve. However, challenges such as ensuring generalization across languages, improving performance in noisy environments, and reducing latency in real-time applications remain critical.

The survey of 12 papers highlights the diverse methodologies used to address these challenges, offering insights into the current state of ASR and RAG technologies. By categorizing the papers into key themes such as multilingual ASR, noise robustness, real-time transcription, and RAG, this work underscores the potential of hybrid architectures that combine retrieval and generation mechanisms with noise-resilient ASR systems.

Future research should focus on improving performance for low-resource languages, advancing adaptive noise-reduction techniques, and enhancing the scalability of RAG models for large-scale applications. Reducing hallucination in generative models and improving fact verification processes will be crucial for the broader adoption of these systems in high-stakes environments.

REFERENCES

[1] V. Karpukhin, T. Sainath, R. Weiss, B. Li, P. Moreno, and E. Weinstein, “Multilingual Speech Recognition With A Single End-To-End Model,” in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An Overview of Noise-Robust Automatic Speech Recognition,” IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 22,no. 4, pp. 745–770, Apr. 2014.

[3] B. Zheng, J. Hu, G. Zhang, Y. Wu, and J. Deng, “Analysis of Noise Reduction Techniques in Speech Recognition,” in Proc. ITNEC, 2020, pp. 1624–1631.

[4] R.Ranchal,T.Taber-Doughty,Y.Guo,K.Bain,H.Martin,J. P. Robinson, and B. S. Duerstock, “Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom,” IEEE Trans. Learning Technol., vol. 7, no. 4, pp. 367–377, Dec. 2014.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey,and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” arXiv preprint arXiv:1711.01694, 2023.

[6] C.Arriaga,A.Pozo,J.Conde,andA.Alonso,“Evaluation of Real-Time Transcriptions Using End-to-End ASR Models,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2024, pp. 3401–3408.

[7] W.KangandD.Roy,“PromptingLargeLanguageModelswith Audio for General-Purpose Speech Summarization,” in Proc. Int. Conf. Spoken Language Processing (ICSLP), 2024.

[8] F.J.Gonzalez-Dominguez,D.Eustis,I.Lopez-Moreno,A.Senior,Beaufays,andP.J.Moreno,“AReal-TimeEnd-to-EndMulti-lingualSpeechRecognitionArchitecture,”inProc.Interspeech, 2014, pp. 2063–2067.

[9] V.Karpukhin,B.Og˘uz,S.Min,P.Lewis,L.Wu, S.Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open- Domain Question Answering,” in Proc. 58th Annual MeetingoftheAssociationforComputationalLinguistics(ACL),2020,pp.6767–6783.

[10] R. C. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. Ø. Rasmussen, C. Matuszek, and B. S. Alexandrov, “Domain-SpecificRetrieval-AugmentedGenerationUsingVectorStores, Knowledge Graphs, and Tensor Factorization,” J. Machine Learning Res., 2023.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Ku¨ttler, M. Lewis, W. Yih, T. Rockta¨schel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-IntensiveNLPTasks,”inAdvancesinNeuralInfor- mation Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.

[12] J. Gao, H. Sun, C. Cao, and Z. Du, “Human Transcription Quality Improvement,” in Proc. IEEE Conf. Spoken Language Technol. (SLT), 2024, pp. 415–422.