IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

A Review of AI-Enhanced Speech Recognition Systems

Prajwal Andanur[1], Suraj S Airani[2], Chandan N R[3], Kiran M S[4], Dr. Harish Bhat[5]
Under Graduate Students, Department of Electronics and Communication Engineering[1-4]
Assistant Professor, Department of Electronics and Communication Engineering[5]
Alvas Institute of Engineering and Technology, Mijar, Mangalore, India

Abstract: *Human-computer interaction has been revolutionized by voice recognition, which enables robots to comprehend and accurately record human speech. Accent, sound, and contextual nuances were too much for early systems, which depended on statistical models and rule-based algorithms. The field has reached previously unheard-of levels of precision thanks to AI, especially neural networks like transformers. From the beginning of speech recognition systems to the most recent developments in artificial intelligence (AI), such as deep machine learning, end-to-end training, and transformer-based architecture, this overview focuses on these themes. It also emphasizes how crucial various datasets like Svarah are for resolving issues with accents and other multilingual situations.Additionally, it explores how Retrieval-Augmented Generation (RAG) might be integrated for contextual understanding and talks about applications in client service, healthcare, education, and accessibility. Data biases, computing needs, and ethical issues still pose significant challenges in spite of these advancements. By working together to innovate solutions, we can fully utilize AI-powered voice recognition to build an interconnected inclusive society.*

Keywords: Speech Recognition, Artificial Intelligence (AI), Deep Learning, Transformer Architectures, Retrieval-Augmented Generation (RAG), Multilingual Contexts, Svarah Dataset, Human-Computer Interaction, End-to-End Learning, Accessibility Technologies

I. INTRODUCTION

Fundamentals of Speech Recognition

Communication with digital systems has become easier and more effective due to speech recognition technology, which has completely transformed human-machine interactions. Enabling robots to comprehend and reliably transcribe spoken language into text is the primary objective of this discipline, which helps close communication barriers and opens up a variety of applications across sectors.

Statistical models and rule-based algorithms played a major role in early voice recognition systems. The intricacy, uncertainty, and contextual subtleties of spoken language often proved too much for these methods to manage, leading to poor accuracy and relevance. An important development in this line of work has been the development of machine learning and artificial intelligence (AI). The accuracy and versatility of contemporary voice recognition systems have grown substantially thanks to the use of neural networks, deep learning, and sophisticated algorithms.

These days, these systems provide excellent outcomes by regulating background noise, comprehending a variety of dialects, and analysing context. Numerous developments, such as automated transcription services, language translation tools, and virtual assistants like Siri and Alexa, have been made possible by this paradigm shift. Robust architectures such as transformers, convolutional neural networks (CNNs), and recurrent neural networks (RNNs)are traits found in AI-enhanced voice recognition systems. Sequential audio data maybe processed using these models, which capture temporal relationships and identify patterns with amazing precision. Additionally, accuracy has been enhanced by integrating self-supervised learning methods with big datasets, enabling models to generalize well across languages, accents, and noisy situations.

This study analyzes the development of voice recognition systems, focusing on major developments made possible by artificial intelligence. We examine well-known models, the technologies that enable them, and practical uses. We also look at how databases like Svarah can improve these systems by guaranteeing accuracy and broadening applicability in

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

a variety of language situations. We seek to offer a thorough grasp of the revolutionary potential of AI-driven speech recognition technology by confronting the opportunities and difficulties in this quickly developing sector.

The Evolution of Speech Recognition

The evolution of voice recognition technology has been an incredible journey, featuring paradigm shifts and ground-breaking discoveries. Over the years, engineers and researchers have gone through a number of creative phases that have all improved the complexity and effectiveness of modern systems.

Rule-Based Frameworks

Rule-based systems were used in the first voice recognition experiments. Due to their foundation in previously determined grammatical and phonetic standards, these early systems were heavily reliant on structured input and specialized programming. Given this, these systems were able to operate in restricted settings, such as digit recognition or simple command inputs, and had narrow vocabularies. Systems could identify spoken numbers or basic words like "start" or "stop," for illustration. Although these methods demonstrated how automated speech recognition was feasible, their applicability in real-world situations was severely limited by their inability to handle variations in pronunciation, accents, and background noise.

Models of Statistics

A significant improvement was made in the 1970s and 1980s with the introduction of statistical modelling. During this time, Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) developed as the mainstays of speech recognition systems. These statistical techniques, as opposed to rule-based ones, introduced the idea of probabilistic reasoning, which allowed systems to analyse voice data for patterns.

While GMMs aided in capturing acoustic variability by describing sound distributions, HMMs concentrated on the temporal variability of speech. Expanded vocabularies and more flexible systems were made possible by such ideas. For example, the practical value of statistical voice recognition was demonstrated by the emergence of applications like automated call routing and simple dictation software. However, these methods had limitations, frequently requiring significant manual feature engineering. Specialists painstakingly extracted traits including spectral features, formants, and pitch. Additionally, these systems performed poorly in noisy environments or with non-standard accents, highlighting the need for more advanced techniques.

The Modern Age of Deep Learning

Speech recognition witnessed an enormous transformation with the advent of deep neural networks in the early 2010s. Neural networks, especially Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and more recently, Transformer models, revolutionized how machines understand and receive speech. These models, contrary to their predecessors, require no human feature extraction since they automatically extract pertinent characteristics from raw data.

Since speech is sequential, RNNs and LSTMs performed extremely well in addressing temporal dependencies and comprehending context. However, they encountered challenges like vanishing gradients, which hindered their ability to effectively navigate lengthy sequences. Transformers, which use self-attention processes to analyse large sequences at once, resolved this issue. Transformers have raised the bar for accuracy and adaptability, as illustrated by models such as Whisper AI and OpenAI's GPT-based configurations. Applications in fields like real-time transcription, virtual assistants, and accessibility aids are made possible by their exceptional capacity to handle a variety of accents, languages, and noisy inputs.

Significant Breakthrough Points in the Evolution of Speech Recognition

1960s: IBM's Shoebox gadget demonstrated that it was capable of recognizing numbers and arithmetic operations.
1980s: The foundation for statistical models was set up by DARPA's Speech Understanding Research (SUR) program.
2000s: Dictation software acquired popularity thanks to commercial programs like Dragon NaturallySpeaking.
2010s: AI-powered speech interfaces were added to everyday electronics by Siri, Google Assistant, and Alexa.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

AI's Contribution to Current Speech Recognition

Speech recognition has become a revolution thanks to AI-powered systems, which have made previously impossible breakthroughs possible. These developments have changed how we use technology by increasing the accessibility, effectiveness, and versatility of voice recognition.

Automated Speech Recognition (ASR)

OpenAI's Whisper and other AI models have raised the bar for voice recognition precision. These systems are dependable and inclusive since they excel at managing a wide range of accents, dialects, and languages. Transformer-based architectures allow models like Whisper to attain remarkable accuracy while requiring little fine-tuning to adjust to non-standard speech patterns. Applications that span multilingual areas, customer service centers, and accessibility technologies depend heavily on this flexibility.

Language Models

In recent voice recognition systems, sophisticated language models like GPT and Llama3.x are essential. By strengthening the contextual comprehension of voice inputs, these models make it possible to provide responses that are both pertinent and cohesive. They may successfully comprehend complex or confusing voice inputs due to their ability to process vast amounts of data and identify intricate patterns. In conversational AI, where interpretation of subtleties and intent is crucial, this skill is especially beneficial.

End-to-End Learning

In the past, a lot of preprocessing and feature extraction methods were frequently used in voice recognition systems. On the other hand, raw audio data may be used to directly train AI-driven end-to-end learning models. This method enhances the system's overall efficiency, streamlines the development process, and minimizes potential sources of error. Comprehensive audio data processing underpins this approach.

Challenges in Speech Recognition for Non-Native Accents

Speech recognition systems encounter numerous challenges when processing non-native accents, particularly in multilingual and diverse countries like India. These challenges are a result of:

Variability in Pronunciation

Pronunciation varies widely among non-native speakers due to the influence of their native languages. This variability can include differences in phoneme articulation, stress patterns, and rhythm.

Lack of Diverse Datasets

Many existing speech datasets primarily focus on native accents or specific regional accents, resulting in a lack of representation for non-native speakers. This scarcity hinders the ability of automatic speech recognition (ASR) systems to generalize effectively across diverse accents.

Underrepresented Languages and Dialects

Regional dialects and languages with smaller speaking populations frequently lack adequate linguistic resources, including lexicons, phonetic information, and annotated recordings of speech.

Code-Switching

In nations with many languages, people frequently switch between them in a single declaration or exchange of words. ASR systems trained on monolingual data have significant challenges from this phenomenon, which is referred to as code-switching.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

Auditory and Ambient Variables

Accurate transcription is made more difficult by differences in recording quality, background noise, and speaker-specific characteristics like pitch and speaking tempo. In non-native contexts, where recording conditions might not be consistent, these variables are more noticeable.

Researchers have underlined the need for datasets that adequately reflect the linguistic and cultural variety of non-native speakers to solve these issues. Furthermore, to enhance ASR performance in such circumstances, developments in phoneme adaptation and robust language modelling are crucial.

The Dataset of Svarah

Resolving accent-related issues in voice recognition, especially for Indian English accents, has advanced greatly with the Svarah dataset. This dataset was created with an emphasis on authenticity and diversity with the express purpose of improving ASR system training and assessment in a multicultural environment.

Vital Components of the Svarah Dataset

Diversity

India is the second largest English-speaking country in the world with a speaker base of roughly 130 million. Unfortunately, Indian speakers find a very poor representation in existing English ASR benchmarks such as LibriSpeech, Switchboard, Speech Accent Archive, etc. We address this gap by creating Svarah, a benchmark that contains 9.6 hours of transcribed English audio from 117 speakers across 65 districts across 19 states in India, resulting in a diverse range of accents.

Realism

Natural conversation fluctuations that arise in daily communication are captured by Svarah. It has regional impacts in prosody, phoneme substitutions, and distinctive intonation patterns. The dataset is especially useful for training ASR models that have the goal of working effectively in real-world situations because of its realism.

Multilingual Context

Although Svarah's main objective is Indian English, it also takes note of linguistic processes like code-mixing, in which English speech is influenced by aspects of native languages. It is particularly helpful in creating ASR systems for multilingual situations due to this characteristic.

Leveraging the Svarah Dataset towards Applications

Model Training

The dataset is an effective instrument for training state-of-the-art ASR models, like Whisper, which need exposure to a variety of syntax patterns in order to generalize well.

Assessment and Benchmarking

People have different accents and varied voices, making it challenging to build an ASR system that understands everyone. It's even more challenging for those who speak multiple languages because their accents are more diverse.Svarah offers a benchmark for assessing ASR performance on Indian English accents. It helps researchers to spot performance gaps in models and improve algorithms to address issues unique to accents.

Multilingual and Ethnic Environments

The Svarah dataset plays a key role in developing ASR systems that serve multilingual populations because of its concentration on regional and cultural variation. Applications like voice assistants, transcription services, and language learning platforms depend heavily on these systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

Leveraging the Svarah Dataset for AI Models

Introduction

The Svarah dataset demonstrated itself to be a useful tool for improving the functionality of several AI-based Automatic Speech Recognition (ASR) systems, especially when it comes to handling the subtleties of Indian accents. This dataset assisted two well-known AI models—OpenAI's Whisper and NVIDIA NeMo—achieve major improvements in the robustness and accuracy of voice detection.

Enhancing AI Models with Svarah Dataset

Whisper: Cutting-edge ASR Model

For instance, when refined using the Svarah dataset, Whisper, a flexible and cutting-edge ASR model, showed significant improvements in transcription accuracy for Indian English. This model, which is renowned for its capacity to generalize over a wide range of languages and accents, was successful in overcoming a number of obstacles related to the linguistic variety of India, including variances in pronunciation, tone changes, and the tendency for code-switching that is frequently seen in Indian speakers.

NVIDIA NeMo: Optimizing Accent Handling

The Svarah dataset was additionally utilized by NVIDIA NeMo, a sophisticated framework for creating and optimizing ASR models, to improve its capacity to handle accent variations. The resilience of these algorithms has been greatly increased by fine-tuning using Svarah data, which allows them to analyse and properly transcribe speech from speakers with a variety of regional accents throughout India. This has been essential in developing more effective and inclusive ASR systems that serve more types of users in multilingual and multicultural settings.

Significance of the Svarah Dataset

The Svarah dataset's significance in bridging the gap between regional language requirements and worldwide ASR technology is highlighted by its incorporation into these models. This dataset promotes technological inclusion and innovation by resolving accent-specific issues and helping to create more dependable and accessible voice recognition solutions for Indian users.

Speech Recognition Employing Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG), a technique that combines retrieval-based approaches with generative language models, is one of the most important developments in voice recognition. By allowing models to dynamically collect pertinent data from large databases or knowledge repositories throughout the speech-to-text process, RAG signifies a paradigm leap. In voice recognition applications, this dynamic retrieval capacity has opened up new possibilities for accuracy and contextual comprehension.

Advancements in RAG for Voice Recognition

Overcoming Conventional Constraints

Conventional speech recognition systems frequently only use pre-trained models, which might falter when handling complicated, domain-specific, or dynamic input. RAG overcomes this constraint by adding real-time retrieval techniques to language models' generating capabilities. This combination guarantees that replies are current, contextually appropriate, and syntactically valid.

Applications of RAG

Real-Time Health Advice

Providing real-time health advice is a major use of RAG. For instance, consumers can query a voice recognition system driven by RAG to provide personalized, evidence-based suggestions for treating chronic diseases. By reducing the need on static databases, this approach guarantees that consumers get the most up-to-date and pertinent information.

Customer Service Enhancement

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

RAG has significantly enhanced the functionality of voice assistants and chatbots in customer service. These systems may now deliver thorough, contextually appropriate answers to user questions by retrieving data from knowledge bases particular to the firm. This development lessens the operational load on human support teams while simultaneously increasing customer happiness.

Accessibility for the Visually Impaired

The implementation of RAG to assist the blind and visually impaired is another specialized but significant use. Speech recognition systems can now read texts aloud, describe settings, and offer navigational aid with a degree of accuracy and customization that was previously impossible by utilizing real-time data retrieval. Users receive more accessibility and freedom because of this capacity.

RAG's Impact

RAG's strength is its capacity to close the gap between dynamic inquiry and static knowledge. RAG systems create replies that are extremely relevant to user demands and context-aware by continually integrating data from well-chosen sources. Because of this, the technology is extremely useful in specialized industries where precision and flexibility are critical, such as legal consultations, medical diagnostics, and instructional tools.

With everything considered, RAG has redefined the possibilities and uses of voice recognition, setting a new standard. It offers unmatched precision in replies and allows for greater understanding of user intent by fusing the advantages of retrieval techniques with generative models. The use of RAG in voice recognition is anticipated to grow as research advances, spurring innovation in a variety of fields.

Integration of Speech-to-Text (STT) and Text-to-Speech (TTS)

Text-to-Speech (TTS) and Speech-to-Text (STT) technologies are integrated into modern automated speech recognition (ASR) systems to provide substantial functionality and user experience. Advances in human-computer interaction have been made possible by this collaboration, which has improved the usability, accessibility, and versatility of speech interfaces. Innovations in many fields have been rendered available by the seamless conversion of spoken language into text and vice versa.

Key Advancements

Intelligent Whisper for Accurate Transcription

Even in loud settings, the sophisticated deep learning model Whisper AI is excellent at accurately identifying and recording human voice. It makes use of large datasets to ensure accurate transcription for a variety of languages, dialects, and accents.

AI-Powered Speech Recognition and Text-to-Speech (TTS) Technologies

AI-powered speech recognition and text-to-speech (TTS) technologies are revolutionizing various industries by enabling seamless human-computer interaction. These advancements have become vital tools in sectors such as media subtitling, judicial procedures, and healthcare documentation, which demand high-fidelity transcriptions.

Advancements in Text-to-Speech (TTS) Systems

New TTS Systems for Voices That Sound Natural

Contemporary TTS systems leverage neural networks such as Tacotron and WaveNet to generate speech that is nearly indistinguishable from human voices. These technologies support inclusiveness and customization by accommodating a variety of languages and accents. Developers can also modify TTS outputs to suit diverse user preferences, including variations in intonation, speed, and tone.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

Applications of TTS and STT Technologies

Voice Bots for Individuals with Visual Impairments

Voice bots enhance daily activities such as navigating applications, reading documents, and shopping online by providing real-time aural feedback and guidance.

Assistants for Voice with Several Spoken Languages

Multilingual voice assistants like Google Assistant and Alexa promote accessibility by breaking language barriers. These systems enable users to communicate in their preferred language, fostering inclusivity across smart home automation, customer service, and education applications.

Emerging Applications

The integration of TTS and STT technologies has paved the way for innovative applications such as voice-activated robots, real-time language translation, and accessibility solutions for individuals with hearing and speech impairments. By advancing these technologies, developers are bridging the gap between people and machines, creating solutions that address diverse consumer needs.

Challenges in AI-Powered Speech Recognition

Despite significant progress, several challenges hinder the broader adoption and effectiveness of voice recognition technologies:

1. Information Lack for Finished Speech

The lack of comprehensive datasets that reflect diverse regional dialects and global accents remains a significant challenge. While datasets like Svarah have partially addressed these gaps, accented speech continues to reduce accuracy and usability for non-native English speakers and those with strong regional accents.

2. High Computational Costs

Advanced AI models like OpenAI's Whisper require substantial computational resources for training and deployment. These high demands disproportionately affect smaller businesses, startups, and researchers with limited resources, raising concerns about sustainability and environmental impact. Developing more energy-efficient ASR (Automatic Speech Recognition) techniques is essential.

3. Security and Legal Issues

The ethical implications of voice recognition technology are another critical concern. Collecting sensitive personal information through speech data raises issues of security, data ownership, and user consent. Biases in training data can result in unfair outcomes, such as reduced identification accuracy for underrepresented groups. Addressing these concerns requires robust data governance practices, anonymization techniques, and transparency in data collection and usage.

Prospects for ASR Systems

The future of ASR technologies involves innovative solutions prioritizing inclusivity, efficiency, and user experience:

1. Fine-Tuning AI Models

Continuous improvement of AI models using diverse datasets, such as Svarah, is crucial. Fine-tuning helps models identify regional accents and subtle speech patterns, closing the accuracy gap for less common linguistic inputs. Transfer learning techniques can also enhance accessibility for underrepresented linguistic groups by adapting existing models with minimal labelled data.

2. Development of Low-Resource Language Models

Creating lightweight ASR systems that perform well with limited computational power is a significant area of focus. Techniques like quantization, pruning, and knowledge distillation are being employed to reduce model size and computational requirements while maintaining accuracy. These advancements could enable the widespread use of ASR technologies in resource-constrained environments, such as rural areas and developing markets.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

3. Multi-Modal AI

Combining multiple data modalities, such as text and visual cues, with voice inputs offers an exciting avenue for enhancing user interactions. For instance, lip-reading technology paired with audio detection can improve accuracy in noisy environments. Similarly, contextual information from visual or textual data can help ASR systems interpret ambiguous sentences. This multi-modal approach has the potential to revolutionize human-AI interactions by making responses more intuitive and adaptable.

4. Development of Real-World Applications

ASR technologies are expanding beyond traditional applications like transcription and virtual assistants. Emerging use cases include real-time translation, accessibility aids for individuals with disabilities, and interaction with smart devices and autonomous vehicles. Continued research, collaboration, and investment in cutting-edge technologies are necessary to ensure these systems are reliable and inclusive.

Use Cases of AI-Powered Speech Recognition

AI-driven speech recognition technologies have transformed numerous industries by enabling seamless human-computer interaction. Key sectors benefiting from these advancements include:

1. Healthcare

Speech recognition technologies are revolutionizing healthcare by providing voice-based diagnostic tools and enabling hands-free interactions with medical devices. These tools enhance clinical documentation, reduce administrative burdens, and improve patient engagement.

2. Education

AI-powered voice recognition is transforming classrooms by serving as interactive AI tutors, offering real-time feedback on grammar and pronunciation. These tools also provide real-time transcripts and captions for students with hearing impairments, fostering inclusivity and personalized learning experiences.

3. Customer Service

Speech models are used in automated customer service lines to handle simple queries and tasks, or to streamline the process of directing the customer to the appropriate human agent. These technologies can escalate issues to human agents when necessary or provide personalized solutions by analyzing customer sentiment and intent. Speech analytics also offer valuable insights for improving products and services.

4. Retail and E-Commerce

Voice-activated shopping assistants enable users to place orders, search for products, and track deliveries using simple voice commands. This technology enhances accessibility and convenience, particularly for individuals with physical disabilities.

II. CONCLUSION

The field of speech recognition has advanced significantly, evolving from basic rule-based systems to sophisticated AI-driven models capable of understanding diverse voices, dialects, and contextual nuances. With datasets like Svarah and technologies such as Retrieval-Augmented Generation (RAG), these systems have achieved remarkable improvements in reliability and contextual awareness.

Despite these advancements, challenges such as ethical concerns, computational costs, and data biases must be addressed to gain widespread trust and acceptance. Continuous research and collaboration are essential for reducing resource consumption and enhancing system inclusivity.

AI-powered speech recognition holds immense potential to transform human-computer interaction and promote accessibility and inclusivity. By fostering multidisciplinary collaboration and ethical innovation, the field can unlock new possibilities for a truly intelligent and connected society.

REFERENCES

[1]. ResearchGate.
Speech Recognition System: A Review.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

[2]. WE ARE SHAIP.
Choosing the Right Speech Recognition Dataset for Your AI Model.
[3]. Symbl.ai.
A Guide to Building an End-to-End Speech Recognition Model.
[4]. ARXIV.
A Review of Deep Learning Techniques for Speech Processing.
[5]. Label Your Data.
Automatic Speech Recognition: Best Practices & Data Sources.
[6]. OpenAI.
Robust Speech Recognition via Large-Scale Weak Supervision.
[7]. Picovoice.
Open-source Speech-to-Text Datasets.
[8]. ResearchGate.
Automatic Speech Recognition: A Review.
[9]. The Gradient.
New Datasets to Democratize Speech Recognition Technology.
[10]. MDPI.
A Speech Recognition Method Based on Domain-Specific Datasets.
[11]. Meta AI.
Wav2vec: State-of-the-art Speech Recognition through Self-Supervision.
[12]. IJFMR.
Automatic Speech Recognition Through Artificial Intelligence.
[13]. Intel Communities.
Enhance Automatic Speech Recognition with Intel AI Solutions.
[14]. ASHA Publications.
Automatic Speech Recognition of Conversational Speech in Individuals with Speech Disorders.
[15]. ARXIV. (Javed, T., Joshi, S., Nagarajan, V., Sundaresan, S., Nawale, J., Raman, A., Bhogale, K., Kumar, P., & Khapra, M. M.). *Svarah: Evaluating English ASR Systems on Indian Accents*.
[16]. Google AI Blog. End-to-End Models for Speech Recognition and Translation.
[17]. Microsoft Research. Speech Recognition Using Transformer Networks.
[18]. NVIDIA Developer Blog. Using AI for Improved Speech Recognition Accuracy.
[19]. Speech Technology Magazine. Advancements in Speech Recognition for Multilingual Contexts.
[20]. IEEE Transactions on Audio, Speech, and Language Processing. Deep Learning Approaches for Speech Processing.
[21]. Nature AI. Transformer-Based Architectures in Speech Recognition.
[22]. SpringerLink. Speech-to-Text Systems: Recent Trends and Future Directions.
[23]. ScienceDirect. Speech Recognition Technologies for Accessibility and Healthcare.
[24]. ACM Digital Library. End-to-End Learning in Modern ASR Systems.
[25]. Hinton, G. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. (IEEE Signal Processing Magazine).
[26]. Kaggle Datasets. Speech Commands Dataset for Neural Networks.
[27]. MIT Technology Review. AI-Powered Voice Recognition Systems: Challenges and Opportunities.
[28]. Nature Reviews. Applications of Artificial Intelligence in Speech Recognition.
[29]. Proceedings of ICASSP. Large-Scale Data for AI Speech Recognition Models.
[30]. GitHub. Open-Source Projects for Speech Recognition.
[31]. DeepMind. Advances in Robust Speech Recognition Models.
[32]. Oxford Academic. Evolution of Neural Architectures in Speech Recognition.
[33]. arXiv. Context-Aware Speech Recognition with Neural Networks.
[34]. AI and Society Journal. Ethical Implications of AI-Driven Speech Recognition.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, December 2024

[35]. Google Research. Training Multilingual ASR Models with Code-Switching.
[36]. Springer AI. Cross-Lingual Transfer in Speech Recognition Models.
[37]. Harvard Data Science Review. Speech Recognition's Role in Accessibility.
[38]. Apple Machine Learning Journal. End-to-End Models for On-Device Speech Recognition.
[39]. ACM Transactions on AI. Addressing Accent Variability in ASR Systems.
[40]. AI Magazine. Challenges and Trends in Speech Processing with Transformers.
[41]. Proceedings of NeurIPS. Self-Supervised Pretraining for Speech Recognition.
[42]. Frontiers in Artificial Intelligence. Speech Recognition and Natural Language Understanding.
[43]. PLOS ONE. Building Resilient ASR for Multilingual Populations.
[44]. IBM Research Blog. Historical Perspectives on Speech Recognition.
[45]. AI Ethics Journal. Managing Bias in AI-Powered Voice Assistants.