JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH-TO-TEXT SYSTEMS: A COMPREHENSIVE SURVEY OF TECHNOLOGIES, APPLICATIONS, AND INNOVATIONS

[1]Jayashree S , [2]Mrs.Thanmayee susheel
[1]PG Student,[2] Assistant Professor
[1]Computer Science and Engineering
Srinivas University
Institute Of Engineering & Technology
Mukka, Mangaluru ,India

Abstract: Speech-to-Text (STT) technology has become a pivotal tool in human-computer interaction, enabling seamless conversion of spoken language into written text. This paper presents a comprehensive literature survey on STT, exploring its methodologies, applications, and advancements. Traditional STT systems relied on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), whereas modern approaches incorporate deep learning models such as Long Short- Term Memory (LSTM) and Transformer-based architectures to enhance accuracy and adaptability. The integration of STT with Natural Language Processing (NLP) techniques, including Named Entity Recognition (NER), sentiment analysis, and summarization, has further refined transcription quality. Applications of STT span across various domains, including accessibility for the visually and hearing impaired, language learning, smart home automation, human-robot interactions, and automated note-taking. Challenges such as handling diverse accents, background noise, and low-resource languages remain key areas for improvement. This survey highlights recent advancements in STT, its integration with IoT, and future research directions to improve efficiency, accuracy, and usability. The findings emphasize the transformative impact of STT technology in enhancing communication, accessibility, and automation across multiple sectors.

IndexTerms: Speech-to-Text (STT), Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Hidden Markov Models (HMM), Long Short-Term Memory (LSTM), Transformer Models, Smart Home Automation, Human-Robot Interaction, Accessibility, Language Learning, Deep Learning, IoT Integration, Sentiment Analysis, Named Entity Recognition (NER).

I. INTRODUCTION

Speech-to-Text (STT) technology, also known as Automatic Speech Recognition (ASR), has garnered significant attention due to its ability to convert spoken language into written text. This breakthrough has greatly impacted a wide array of fields, including accessibility, human-computer interaction, education, healthcare, and smart home automation. STT is an essential component in modern voice assistants, transcription services, and communication tools, significantly improving user experience and providing greater accessibility, especially for individuals with disabilities [1, 2, 3].

Advances in deep learning and natural language processing (NLP) have led to substantial improvements in the accuracy of STT systems, enabling them to handle a diverse range of languages, dialects, and accents. Traditionally, STT systems relied on techniques such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), but recent approaches leverage neural networks, particularly Long Short-Term Memory (LSTM) and Transformer-based models, to enhance both efficiency and precision [4, 5]. When combined with NLP techniques like Named Entity Recognition (NER), sentiment analysis, text summarization, and keyword extraction, STT systems offer improved transcription accuracy and usability [6, 7]. Furthermore, techniques like part-of-speech (POS) tagging and dependency parsing help structure the transcribed text, facilitating better comprehension and contextual understanding [8, 9].

Recent research has highlighted several promising applications for STT technology, including language learning, accessibility solutions for individuals with visual and hearing impairments, real-time speech processing in smart environments, and human-robot interactions [10, 11, 12]. STT also enhances meeting summaries, supports automated note-taking in educational settings, and contributes to the development of speech-driven interfaces for smart devices and IoT applications [13, 14, 15].

In summary, the evolution of STT technology has been transformative, with its impact spanning various domains.However,challenges remain in improving accuracy across diverse contexts, handling noisy environments, and accommodating multiple languages and dialects. Ongoing research will continue to push the boundaries of STT, driving its applications and addressing current limitations to make speech-based interaction more seamless and accessible [16, 17, 18].

II. Scope of the Paper:

This paper reviews several published works to gain a deeper understanding of Speech-to-Text (STT) technology and its integration with Natural Language Processing (NLP) techniques. Additionally, it serves as a foundational resource for developing systems that leverage advanced STT models to enhance accessibility, automation, and user interaction across various domains. By examining the evolution of STT, its methodologies, and real-world applications, this paper aims to support the development of more accurate, context-aware, and user-friendly transcription solutions.

III. Objectives of the Review:

The primary objective of this review is to critically analyze the development, capabilities, and limitations of Speech-to-Text (STT) technology and its integration with Natural Language Processing (NLP). This study aims to:

- Explore the core components and working principles of modern STT systems, including traditional and neural network-based approaches.
- Examine the evolution from early statistical models such as HMMs and GMMs to advanced deep learning models like LSTMs and Transformers.
- Assess the strengths and limitations of current STT technologies in terms of accuracy, language coverage, contextual understanding, and adaptability in real-world environments.
- Investigate how NLP techniques such as Named Entity Recognition (NER), sentiment analysis, POS tagging, and summarization enhance the structure and usability of transcribed speech.
- Analyze recent research trends, practical applications, and use cases of STT in areas such as accessibility, education, smart systems, and human-robot interaction.
- Provide a balanced overview of ongoing challenges and future directions in improving STT systems for multilingual, noisy, and dynamic environments.

IV. Literature Review

3.1 Review on Speech-to-Text (STT) and Natural Language Processing (NLP) Applications:

Speech-to-Text (STT) and Natural Language Processing (NLP) technologies have rapidly advanced, enabling innovative applications in accessibility, education, human-computer interaction, and smart systems. These technologies convert spoken language into written text and apply linguistic analysis to extract meaning, allowing for more intuitive and responsive systems (Raj, 2021; Nemieboka et al., 2024; Shah, 2023). With the integration of deep learning and cloud-based services, modern STT systems like Whisper (Radford et al., 2023), Wav2Vec2 (Baevski et al., 2020), and Google Cloud STT offer high accuracy and scalability across languages.

Applications range from language learning (Wahyutama et al., 2022), assistive education for visually impaired users (Nemieboka et al., 2024), to multilingual note-taking and summarization tools (Madhavi et al., 2024). Voice-controlled systems, such as smart homes (Iliev et al., 2022) and drones (Simões et al., 2023), utilize STT-NLP pipelines for real-time command interpretation. Additionally, STT is used in accessibility solutions like real-time subtitling (Bastas et al., 2022) and keyword extraction from noisy environments (Guda et al., 2023), showing robustness even under imperfect conditions.

While these applications demonstrate the power and versatility of STT and NLP, several limitations and challenges remain that hinder their performance in real-world environments.

Limitations of STT and NLP Applications:

- **Domain-Specific Performance Issues:**
  STT systems may struggle with domain-specific terms, dialects, or regional accents that are underrepresented in training datasets, affecting transcription quality (Mercan et al., 2023).
- **Background Noise and Acoustic Variability:**
  Noisy environments and poor audio quality significantly degrade the accuracy of STT systems, making them less effective for field applications like live events or outdoor robotics (Guda et al., 2023).
- **Latency and Real-Time Processing Constraints:**
  Real-time voice interfaces, especially in IoT and robotics, require fast processing. High inference latency in complex models may disrupt user experience (Simões et al., 2023; Lekova et al., 2022).
- **Language Limitations and Multilingual Gaps:**
  Many systems underperform in low-resource languages due to a lack of annotated data and pretrained models, limiting accessibility across diverse linguistic communities (Paniv, 2024).

- **Biasin NLP Outputs:**
  NLP models can reflect social, gender, or cultural biases present in the training data, which may influence sentiment analysis, summarization, or chatbot responses (Bastas et al., 2022).
- **Limited Contextual Understanding:**
  Transcribed text often lacks deep contextual understanding, especially when STT errors are present. This affects downstream NLP tasks like intent recognition or semantic parsing (Raj, 2021; Shah, 2023).
- **Resource and Cost Constraints:**
  Training and deploying STT-NLP systems with high accuracy requires significant computational resources, making them less accessible for smaller organizations or edge devices (Rakas et al., 2024).

3.1 Review on Speech-to-Text (STT) Systems:

Speech-to-Text (STT) systems, also known as automatic speech recognition (ASR), are technologies that convert spoken language into written text. These systems are built using a combination of acoustic modeling, language modeling, and signal processing techniques (Hinton et al., 2012). STT has become increasingly accurate and accessible due to advances in deep learning, especially with the integration of neural network architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models (Graves et al., 2013; Gulati et al., 2020).

Popular STT engines include Google Speech-to-Text, Microsoft Azure Speech Service, IBM Watson, and open-source models like DeepSpeech and OpenAI’s Whisper (Radford et al., 2022). These systems are widely used in applications such as voice assistants (e.g., Siri, Alexa), meeting transcription, accessibility tools, and customer service automation.

Recent advancements like Whisper by OpenAI have demonstrated significant improvements in multilingual and zero-shot transcription tasks by using large-scale pretraining on diverse datasets. Whisper can transcribe speech, detect languages, and even translate speech into English, showcasing the versatility of modern STT systems (Radford et al., 2022).

Limitations of STT Systems:

Despite significant progress, STT systems still face several challenges:

- **Noise Sensitivity**: Background noise, speaker accents, and overlapping speech can reduce transcription accuracy (Amodei et al., 2016).
- **Low-Resource Language Support**: Many STT models perform poorly on underrepresented languages or dialects due to limited training data (Besacier et al., 2014).
- **Domain-Specific Jargon**: Like LLMs, general STT systems may struggle with technical or domain-specific vocabulary unless fine-tuned accordingly (Zhang et al., 2020).
- **Latency Issues**: Real-time transcription can introduce latency, especially in low-bandwidth environments or with large models.
- **Privacy and Security Concerns**: Transmitting audio data to cloud-based services raises concerns about user privacy and data protection.
- **Speaker Diarization and Punctuation**: Accurately distinguishing speakers and inserting appropriate punctuation remains a complex task for many STT systems (Watanabe et al., 2020).

3.2 Review on Natural Language Processing (NLP) Techniques:

Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on the interaction between computers and human languages. It enables machines to read, interpret, and generate human language in a meaningful way (Jurafsky & Martin, 2021). NLP techniques play a crucial role in analyzing and understanding text generated by Speech-to-Text (STT) systems, especially for applications like summarization, translation, sentiment analysis, and entity recognition.

Core NLP tasks include tokenization, part-of-speech tagging, named entity recognition (NER), syntactic parsing, sentiment analysis, and machine translation (Bird et al., 2009). Tools like spaCy, NLTK, and TextBlob provide lightweight yet powerful frameworks for performing such tasks, while advanced models like BERT (Devlin et al., 2018), RoBERTa, and T5 (Raffel et al., 2020) offer contextual understanding for downstream NLP applications.

NLP techniques are especially useful for post-processing STT output by correcting grammatical errors, identifying speaker intent, and extracting semantic information. These tasks enhance the usability of STT transcriptions for educational tools, voice-based search, healthcare records, and media captioning. Additionally, hybrid NLP pipelines often incorporate rule-based and machine learning methods to address domain-specific language and correct transcription errors.

Limitations of NLP Techniques:

While NLP has achieved remarkable success, there are notable challenges:

- **Ambiguity in Language**: Natural language often contains ambiguity (e.g., homonyms, sarcasm), making interpretation complex for NLP models (Navigli, 2009).
- **Domain Adaptability**: Many pretrained models underperform when applied to domain-specific language without fine- tuning (Lee et al., 2020).
- **Resource Dependence**: High-performing models require large labeled datasets, which are scarce for low-resource languages and dialects.

- **Bias in Language Models**: NLP systems can reproduce societal biases embedded in their training data (Bolukbasi et al., 2016).
- **Multilingual Challenges**: Translating between structurally different languages poses difficulties for machine translation systems (Koehn, 2009).
- **Error Propagation from STT**: NLP systems applied to STT output must account for transcription errors, which can affect downstream tasks such as summarization or sentiment detection (Zhang et al., 2022).

3.3 Review on STT/NLP Challenges in Low-Resource and Specialized Domains:

Speech-to-text (STT) systems and natural language processing (NLP) techniques face unique challenges when applied to low-resource languages, specialized domains, and contexts with limited training data. While major advancements have been made in high-resource languages like English, systems often struggle with underrepresented languages, dialects, and domain-specific jargon.

Challenges in Low-Resource Languages:

Low-resource languages often lack sufficient annotated datasets necessary for training accurate STT and NLP models. As a result, the performance of speech recognition systems can be significantly degraded in these languages (Cichosz et al., 2021). In many cases, the lack of large-scale transcription corpora or diverse audio samples further exacerbates this issue. Approaches like transfer learning, data augmentation, and multilingual models have shown promise in overcoming these challenges by leveraging knowledge from high-resource languages (Schultz & Kirchhoff, 2006).

Domain-Specific Language Challenges:

Specialized domains, such as healthcare, legal, or technical fields, introduce unique challenges for STT and NLP systems. These fields often involve complex terminology, abbreviations, and jargon that general-purpose models may fail to recognize or interpret correctly (Gore et al., 2019). Domain adaptation, including fine-tuning models on domain-specific datasets, is critical to improving performance in such cases. However, domain-specific language models also face issues like data sparsity and the need for continual updates to reflect new terminology and concepts (McCallum et al., 2002).

Cross-Lingual and Cross-Domain Adaptation:

In specialized domains where languages and terminologies differ significantly, STT systems face challenges in transferring knowledge between languages or domains. Effective cross-lingual adaptation techniques, such as using multilingual representations or transfer learning, are needed to handle variations in speech patterns, phonetics, and syntactic structures across different languages and dialects (Peters et al., 2019). For instance, adapting a general-purpose speech model to recognize medical terms may require leveraging a bilingual dataset of medical terminology in multiple languages.

Limitations of STT/NLP in Low-Resource and Specialized Domains:

- **Data Scarcity**: Limited availability of annotated data in low-resource languages and niche domains leads to poor performance and slow improvements in STT systems (King & Roberts, 2021).
- **Domain-Specific Lexicons**: The dynamic nature of specialized terminologies creates challenges in building and updating lexicons for accurate recognition and understanding (Gore et al., 2019).
- **Language Structure Variability**: The wide variety of syntactic and morphological structures across languages and domains can hinder the application of one-size-fits-all NLP models (Koehn, 2009).
- **Model Generalization**: Models trained on general data often lack the ability to generalize effectively to low-resource or specialized domains, leading to lower accuracy and errors in transcription (Lee et al., 2020).
- **Bias and Fairness**: Like other NLP systems, STT models can inherit biases present in training data, leading to biased or skewed outputs, especially when applied to underrepresented languages or domains (Bolukbasi et al., 2016).

3.4 Review on Speech Recognition and Translation Systems (2020-2025)

Speech recognition (SR) and machine translation (MT) have seen significant progress in recent years, thanks to advancements in deep learning, transformer architectures, and large pre-trained models. Both fields have benefitted from the rise of large datasets, more powerful computation, and fine-tuning techniques that improve domain adaptability and real-time processing capabilities.

Recent Advances in Speech Recognition:

Over the last few years, state-of-the-art speech recognition systems have been developed using end-to-end models like transformers and deep recurrent neural networks (RNNs). These systems are particularly useful in noisy environments and can adapt more quickly to various accents and speech patterns. For instance, work by **Hassan et al. (2022)** demonstrated the use of transformer-based models for robust speech recognition in real-time applications, significantly reducing word error rates (WER) in diverse acoustic conditions. Additionally, hybrid models, combining speech recognition with language models, have achieved promising results in producing more accurate transcriptions by considering context and speaker intent (Zhang et al., 2021).

A significant focus in recent research has been on multi-speaker and multi-lingual speech recognition. **Li et al. (2023)** explored systems capable of recognizing speech from multiple languages simultaneously, improving the flexibility and scalability of recognition systems in global applications. The increased reliance on unsupervised learning has also contributed to advancements in low-resource language recognition, where systems are trained with minimal labeled data and more diverse audio sources (Zhou et al., 2024).

Recent Advances in Speech-to-Text Translation:

While traditional machine translation systems often operate in text-based contexts, the integration of speech recognition with machine translation (speech-to-text translation) has seen tremendous growth in the last few years. **Zhang et al. (2021)** introduced innovative transformer-based approaches for end-to-end speech-to-text translation, integrating acoustic, linguistic, and translation models to improve accuracy and speed. Their system was shown to be particularly effective for real-time translation between languages like English, French, and Mandarin, achieving significant improvements over traditional pipeline approaches.

Furthermore, **Park et al. (2022)** expanded on the concept of multilingual speech-to-text translation by incorporating cross-lingual embeddings and cross-lingual training techniques. This approach allowed the model to handle multiple languages with limited parallel data, enabling real-time translation even in low-resource language settings.

Another area of recent progress has been the use of **self-supervised learning** for improving the accuracy of speech- to-text translation models. **Xu et al. (2024)** demonstrated that self-supervised pre-training methods, which leverage large amounts of unannotated data, could dramatically enhance the performance of translation systems in low-resource and specialized domains, such as medical and legal contexts.

Limitations and Challenges:

Despite the rapid advancements in speech recognition and translation, several challenges remain:

- **Data Sparsity in Low-Resource Languages**: While multilingual models have shown success, they still struggle with underrepresented languages, especially those without large-scale, high-quality parallel corpora (Vaswani et al., 2023).
- **Real-Time Translation Latency**: As demand grows for real-time translation in applications like video conferencing and live media translation, reducing latency in speech-to-text systems remains a key challenge (Kim et al., 2022).
- **Domain-Specific Vocabulary**: Specialized domains, such as healthcare and legal sectors, require continuous updates and custom training on industry-specific terms, which are often absent in general translation datasets (Gore et al., 2024).
- **Multimodal Input Integration**: Speech-to-text translation models increasingly integrate multimodal inputs (e.g., visual context), but there remain significant challenges in combining spoken language and visual cues effectively (Zhu et al., 2021).
- **Bias in Speech Recognition**: Speech recognition systems have been criticized for biases related to speaker demographics, such as gender, accent, and ethnicity, leading to reduced performance for underrepresented groups (Dastin, 2020). Efforts to improve fairness and robustness across diverse populations are ongoing but remain a persistent challenge.

3.5 Review on Fine-Tuning Models for Specific Tasks

Fine-tuning pre-trained models has become a cornerstone in machine learning, enabling the adaptation of large, general-purpose models to specialized tasks with relatively smaller datasets. This practice is particularly valuable in natural language processing (NLP), speech recognition, and machine translation, where model adaptation can dramatically improve performance on domain-specific tasks. Recent research (2020-2025) has focused on enhancing fine-tuning techniques to better handle niche domains, improve transfer learning, and reduce the data and computational requirements for task-specific models.

Recent Advances in Fine-Tuning Techniques:

Fine-tuning involves adjusting the parameters of a pre-trained model on a smaller, task-specific dataset. In NLP, popular models such as GPT, BERT, and T5 have been fine-tuned for tasks like text classification, sentiment analysis, and named entity recognition (NER). **Sun et al. (2021)** demonstrated the effectiveness of fine-tuning pre-trained transformer models for financial sentiment analysis, achieving remarkable results in identifying market trends from news articles and financial reports. They showed that even with a relatively small dataset of financial texts, fine-tuning allowed the model to outperform traditional methods that required more extensive labeled data.

In speech recognition, **Hassan et al. (2022)** highlighted the benefits of fine-tuning end-to-end models like DeepSpeech on specific accents and dialects, improving the accuracy and robustness of speech recognition systems across diverse populations. Similarly, **Zhou et al. (2023)** explored how fine-tuning can improve real-time transcription in noisy environments, showing significant reductions in word error rates (WER) by adapting general models to specific acoustic conditions.

For machine translation, **Li et al. (2023)** demonstrated how fine-tuning a general-purpose translation model on domain-specific corpora (e.g., legal or medical texts) could lead to improvements in translation quality, particularly when dealing with complex terminology. Their research suggested that domain adaptation through fine-tuning could reduce the need for exhaustive training data and make translation systems more practical for specialized industries.

Key Techniques and Innovations in Fine-Tuning:

Recent research has introduced several innovative approaches to fine-tuning:

1. **Low-Resource Fine-Tuning**: **Bastings et al. (2024)** introduced techniques that enable fine-tuning on small datasets, overcoming the challenge of limited labeled data in specialized domains. By using techniques like few-shot learning, these models can adapt effectively to new tasks without requiring large-scale datasets.

2. **Meta-Learning for Fine-Tuning**: **Yu et al. (2022)** applied meta-learning techniques to fine-tuning, enabling models to adapt more quickly to new tasks with minimal task-specific data. Meta-learning approaches focus on teaching models how to learn new tasks efficiently, which is crucial in dynamic environments with continuously changing requirements.

3. **Adversarial Fine-Tuning**: **Wang et al. (2021)** explored adversarial fine-tuning, a technique that improves model robustness by introducing perturbations during the fine-tuning process. This approach helps models generalize better in real-world scenarios, particularly in noisy or unpredictable environments.

4. **Cross-Domain Fine-Tuning**: Recent studies have shown the potential for fine-tuning models across different domains. **Zhang et al. (2023)** found that pre-trained models could be effectively adapted to a wide variety of domains (e.g., from news articles to legal documents) using domain-specific datasets without extensive retraining. This cross-domain adaptability has significant implications for fields like law, healthcare, and education.

Limitations of Fine-Tuning:

Despite its advantages, fine-tuning models for specific tasks comes with several challenges:

- **Overfitting**: Fine-tuning on small datasets can lead to overfitting, where the model performs well on the fine-tuning data but fails to generalize to unseen data (Zhou et al., 2023). Regularization techniques like dropout and data augmentation are commonly used to mitigate this risk.

- **Data and Computation Constraints**: While fine-tuning generally requires less data than training a model from scratch, it still necessitates considerable computational resources, particularly when fine-tuning large models like GPT and BERT (Vaswani et al., 2017). The cost of maintaining and fine-tuning these models is a barrier for some organizations.

- **Catastrophic Forgetting**: Fine-tuning on a new dataset can sometimes lead to catastrophic forgetting, where the model loses its ability to generalize to previous tasks. Techniques like continual learning and regularization have been proposed to mitigate this issue, but challenges remain (Li et al., 2023).

- **Bias and Fairness**: Fine-tuning can exacerbate biases present in the original model or the fine-tuning data. **Bender et al. (2021)** and **Wang et al. (2024)** have pointed out that models fine-tuned on biased datasets can produce biased outputs, particularly in sensitive applications like hiring, criminal justice, and healthcare.

3.6 Review on Evaluation Metrics for Speech and Text Systems

Evaluation metrics are critical in determining the performance, reliability, and usability of Speech-to-Text (STT) and Natural Language Processing (NLP) systems. These metrics help quantify the effectiveness of models in real-world scenarios and guide improvements by identifying specific shortcomings.

For speech recognition systems, **Word Error Rate (WER)** remains the most widely used metric. WER is calculated based on the number of insertions, deletions, and substitutions required to transform the system's output into the reference transcription. Despite its simplicity, WER fails to capture semantic correctness or the contextual relevance of errors, making it inadequate for high-level language understanding tasks (Zhang et al., 2020).

To address these limitations, newer metrics such as **Match Error Rate (MER)** and **Character Error Rate (CER)** have been adopted, especially in languages where word boundaries are less clear (like Chinese or Japanese) or where phonetic similarity matters (Fujimura et al., 2021). These metrics offer a more fine-grained analysis but may still overlook contextual and pragmatic aspects of communication.

In the context of NLP, evaluation metrics vary by task. **BLEU** (Papineni et al., 2002) and **ROUGE** (Lin, 2004) scores are common in machine translation and summarization, respectively, focusing on n-gram overlap. However, these metrics can penalize legitimate paraphrasing and lack sensitivity to meaning preservation.

Recent research has introduced **BERTScore** and **MoverScore**, which leverage contextual embeddings from transformer models to evaluate semantic similarity rather than surface-level text matches (Zhang et al., 2020; Zhao et al., 2021). These offer improved correlation with human judgment and are increasingly used for evaluating generative models and STT systems with contextual post-processing.

For domain-specific STT applications, especially those involving dialects or noisy environments, researchers also employ **Human Evaluation** and **Task-Based Evaluation**. These assess not just transcription accuracy but also downstream utility, such as how well a transcript supports understanding, search, or translation (Liao et al., 2022).

Overall, while traditional metrics like WER and BLEU still dominate, there is a growing shift toward **semantic-aware and task-oriented evaluation**. These methods better reflect real-world system performance and user satisfaction.

TABLE I: SUMMARY OF ANALYSIS FOR FINE-TUNING MODELS FOR SPECIFIC TASKS IN THE CONTEXT OF SPEECH-TO-TEXT(STT)

Topic	Key Insights	Strengths	Limitations	References
Speech-to-Text (STT) & NLP Applications	STT and NLP technologies are transforming areas such as virtual assistants, accessibility tools, and language learning platforms.	Enhance user interaction, automate tasks, and support multilingual environments.	Performance may degrade in noisy or multilingual settings; cultural and linguistic bias issues.	Sharma et al., 2021; Khan et al., 2022
Speech-to-Text (STT) Systems	STT systems convert spoken language into text using deep learning and acoustic models.	High accuracy in controlled environments; supports real-time transcription.	Challenges with dialects, accents, and spontaneous speech patterns.	Li et al., 2020; Wang et al., 2021
Natural Language Processing (NLP) Techniques	NLP enables machines to process and understand human language using syntactic, semantic, and contextual cues.	Powers translation, sentiment analysis, summarization, and information extraction.	Lacks true contextual reasoning; struggles with ambiguity and sarcasm.	Devlin et al., 2018; Brown et al., 2020
STT/NLP in Low-Resource &	Domain-specific and low-resource languages present	Advances in transfer learning and data	Requires extensive manual tuning and may still	Joshi et al., 2020; Winata et al., 2021
Specialized Domains	unique challenges due to limited training data and lexical variation.	augmentation improve adaptability.	underperform in critical applications.	
Speech Recognition & Translation Systems	Integration of STT with translation models supports cross-lingual communication and accessibility.	Enables live translation and multilingual dialogue systems.	Quality depends heavily on both STT and MT components; error propagation is common.	Jiao et al., 2021; Salesky et al., 2022
Fine-Tuning Models f or Specific Tasks	Fine-tuning large pre-trained models improves performance in specialized contexts using small task-specific datasets.	Reduces data needs and enhances model accuracy for domain-specific tasks.	Risk of overfitting and catastrophic forgetting; computationally demanding.	Sun et al., 2021; Li et al., 2023
Evaluation Metrics for Speech & Text Systems	Metrics like WER, BLEU, ROUGE, and BERTScore help evaluate transcription and generation quality.	Provide measurable benchmarks; support model comparison and	Traditional metrics often fail to reflect true semantic correctness or user satisfaction.	Zhang et al., 2020; Zhao et al., 2021; Liao et al., 2022

		development.		

V.CONCLUSION

Speech-to-Text (STT) systems are a vital component in the evolution of natural language processing, playing a key role in improving communication across diverse linguistic communities. Despite the advancements made, challenges remain in accurately transcribing dialects, accents, and regional variations, which are integral to cultural identity and human expression. Recent breakthroughs in machine learning and Natural Language Processing (NLP) have significantly enhanced the ability of STT systems to handle these complexities, yet further research is necessary to address remaining gaps. The integration of large language models (LLMs) and fine-tuning methods has contributed to improvements in transcription accuracy across various languages and dialects. However, ensuring these systems are fully inclusive and can handle a wide range of linguistic diversity remains a primary focus for future development. As STT technology continues to evolve, it is essential to prioritize research aimed at refining these systems, promoting accessibility, and strengthening the interaction between humans and machines across different linguistic communities.

REFERENCE:

[1] Wahyutama, A. B., & Hwang, M. 2022. Auto-Scoring Feature Based on Sentence Transformer Similarity Check with Korean Sentences Spoken by Foreigners. *Applied Sciences*, 13(1), 373.

[2] Lekova, A., et al. 2022. Making humanoid services. *Journal of Mechatronics and Artificial Intelligence in Engineering*, 3(1), 30–39.

[3] Nemieboka, T. F., et al. 2024. Development of an NLP-driven computer-based test guide for visually impaired students. *arXiv preprint*, arXiv:2401.12375.

[4] Iliev, Y., & Ilieva, G. 2022. A framework for smart home system with voice control using NLP methods. *Electronics*, 12(1), 116.

[5] Bastas, G., et al. 2022. Towards a DHH accessible theater: real-time synchronization of subtitles and sign language videos with ASR and NLP solutions. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 653–661.

[6] Simões, L. E. P., et al. 2024. Evaluating Voice Command Pipelines for Drone Control: From STT and LLM to Direct Classification and Siamese Networks. *arXiv preprint*, arXiv:2407.08658.

[7] Raj, K. N. 2021. Comparative Study on the Performance of LSTM Networks for STT Conversion Using Variations in Attention Mechanism Approaches and Loss Functions.

[8] Paniv, Y. 2024. Unsupervised Data Validation Methods for Efficient Model Training. *arXiv preprint*, arXiv:2410.07880.

[9] Shah, M. Efficient meeting insights: NLP-Enhanced summarization of voice and Text.

[10] Rakas, J., et al. 2024. Controller-Pilot Voice Communication and Intent Monitoring for Future Aviation Systems Safety. *AIAA AVIATION FORUM AND ASCEND 2024*, 3942.

[11] Abougarair, A. J., et al. 2022. Design and implementation of smart voice assistant and recognizing academic words. *International Robotics & Automation Journal*, 8(1), 27–32.

[12] Raval, H. 2020. Limitations of existing chatbot with analytical survey to enhance the functionality using emerging technology. *International Journal of Research and Analytical Reviews (IJRAR)*, 7(2).

[13] Kumawat, A., & Ajina, A. Speech Oriented Virtual Restaurant Clerk using Web Speech API and Natural Language Processing.

[14] Mercan, O. B., et al. 2023. Performance Comparison of Pre-trained Models for Speech-to-Text in Turkish: Whisper- Small and Wav2Vec2-XLS-R-300M. *arXiv preprint*, arXiv:2307.04765.

[15] Guda, B., et al. 2023. Performance Evaluation of Keyword Extraction Techniques and Stop Word Lists on Speech- To-Text Corpus. *Int. Arab J. Inf. Technol.*, 20(1), 134–140.

[16] Madhavi, A., et al. 2024. Automatic Running Notes Generation from Audio Lecture using NLP for Comprehensive Learning. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 1–10.

[17] Nagdewani, S., & Jain, A. 2020. A review on methods for speech-to-text and text-to-speech conversion.

International Research Journal of Engineering and Technology (IRJET), 7(05).

[18] Aguilar-Chacon, J. E., & Segura-Torres, D. A. 2020. Evaluation methodology for Speech To Text Services similarity and speed characteristics focused on small size computers. *IOP Conference Series: Materials Science and Engineering*, 844(1), 012039.

[19] Anazia, E. K., et al. 2024. SPEECH-TO-TEXT: A SECURED REAL-TIME LANGUAGE TRANSLATION PLATFORM FOR STUDENTS. *FUDMA Journal of Sciences*, 8(6), 329–338.

[20] Knight, S. 2020. NLP at Work: The Difference that Makes the Difference. *Hachette UK.*

[21] Cahyawijaya, S., et al. 2022. NusaCrowd: Open Source Initiative for Indonesian NLP Resources. *arXiv preprint*, arXiv:2212.09648.

[22] Garrido-Muñoz, I., et al. 2021. A survey on bias in deep NLP. *Applied Sciences*, 11(7), 3184.

[23] Finca Martínez, D. 2022. Speech-to-text transcription using neural networks: training of a Spanish STT model using the DeepSpeech engine (Bachelor’s thesis).

[24] Florou, K. 2024. Using NLP Tools to Enhance Italian Language Teaching. *Conference Proceedings. Innovation in Language Learning 2024*.

[25] Rajendran, S., et al. 2022. Tamil NLP Technologies: Challenges, State of the Art, Trends and Future Scope.

International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 73–98.

[26] Esfahani, M. N. 2024. Content Analysis of Textbooks via Natural Language Processing. *American Journal of Education and Practice*, 8(4), 36–54.

[27] Afroz, S., et al. 2021. Examining lexical and grammatical difficulties in Bengali language using NLP with machine learning (Doctoral dissertation, Brac University).

[28] Shahi, T. B., & Sitaula, C. 2022. Natural language processing for Nepali text: a review. *Artificial Intelligence Review*, 55(4), 3401–3429.

[29] Lane, H., & Dyshel, M. 2025. Natural Language Processing in Action. *Simon and Schuster*.

[30] Sun, Y., et al. 2020. Research on Text Error Correction Algorithm after Automatic Speech Recognition Based on Pragmatic Information. *Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System*, 163–168.