IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multimodal Language Converter

[1]Neeha velagapudi, [2]Sanjana Jatavath, [3]Manaswi Kankanala, [4]Nadia Anjum

[1]Student, [2] Student, [3] Student, [4]Assistant Professor

[1]Department of Artificial Intelligence and Data Science,

[1]Stanley College of Engineering and Technology for Women, Hyderabad, India

Abstract: This paper provides a comprehensive literature survey on major advancements in text and speech processing technologies, such as text–to–speech, speech–to–text, brain-computer interface, cross-lingual translation, and summarization. These technologies have made significant progress in recent years, introducing large language models to the industry for seamless interaction by incorporating techniques that generate apt text to the given prompt by improving accuracy and adapting it to different domains. The inputs are accepted in both text and speech formats for text translation and summarization. This allows the system to process and generate translations and summaries directly from typed or spoken content. For speech-to-text, new adversarial algorithms help improve the model's ability to handle challenging or unpredictable conditions, like noise or distortion in speech. While recent TTs models, like Natural Speech, TextrolSppech, and Style TTs 2. These create natural-sounding voices in languages with limited training data by using style control and VAE (variational Autoencoder) models. Cross-lingual summarization and BCIs are also evolving, with systems like DeWave that translate between languages and from ECG to text, simplifying human-computer interaction. Additionally, Sequence-to-sequence models upgrade speech synthesis and recognition by enhancing their ability to express different ideas in various contexts. Overall, digital signal processing, deep learning, and cross-model learning are leading to the emergence of more flexible, user-friendly, and accurate language technologies for communication.

Index Terms – Introduction, Methodology, Challenges & Limitations, Result Discussion, Conclusion

I. Introduction

Advances in natural language processing (NLP) and speech technologies, which include text-to-speech, speech-to-text, cross-lingual summarization, and brain-computer interfaces, have made human-computer interaction easy [1]. These technologies have a wide range of applications, from virtual assistants to tools for seamless cross-lingual communication. The emergence of large language models and deep learning neural network architectures has driven advancements in expressiveness and robustness. At the same time, models like style TTS and Natural Speech have set new benchmarks for naturalness in TTS across languages [2], [19], even in scenarios where the model has limited training data. Technologies such as adversarial algorithms increase the system's ability to function accurately in noisy environments [6].

Despite the progress and developments, challenges persist such as achieving accuracy in low-resource languages, handling domain-specific vocabulary [3], [10], and minimizing computational requirements. Ensuring robustness across noisy and varied environments also remains essential, especially as these technologies are expected to operate effectively across diverse domains and settings. This survey provides an in-depth overview of contemporary techniques and models in TTS, S2T, cross-lingual summarization, and BCIS. By exploring recent advances, applied methodologies, and ongoing challenges, this paper captures the current research landscape. It suggests directions for future work aimed at enhancing adaptability, precision, and usability in NLP and speech-processing technology.

II. Methodology

The system starts by accepting raw input in the form of either text or audio, such as natural language sentences or spoken words. It then processes this input through a cleaning and normalization phase, utilizing techniques like tokenization for text and Mel-Frequency Cepstral Coefficients (MFCC) for audio. Advanced feature extraction methods, such as Wav2Vec for audio and CNNs for text, help the model grasp essential characteristics.

These features are fed into various models, including LSTMs and Variational Autoencoders, which learn to perform tasks like summarization or speech generation. Depending on the goal, different algorithms are employed for summarization, including extractive and abstractive methods, as well as techniques for enhancing robustness in speech tasks. For applications needing uncertainty annotations, systems like the GPT-4 Annotation System label the data, which is further refined through expert evaluations and iterative self-improvement. The output stage then generates results in the desired format, whether synthesized speech or concise summaries. Lastly, a feedback loop continuously enhances the model's performance by learning from user interactions and its outputs, ensuring it remains effective in real-world applications.

IJCRT

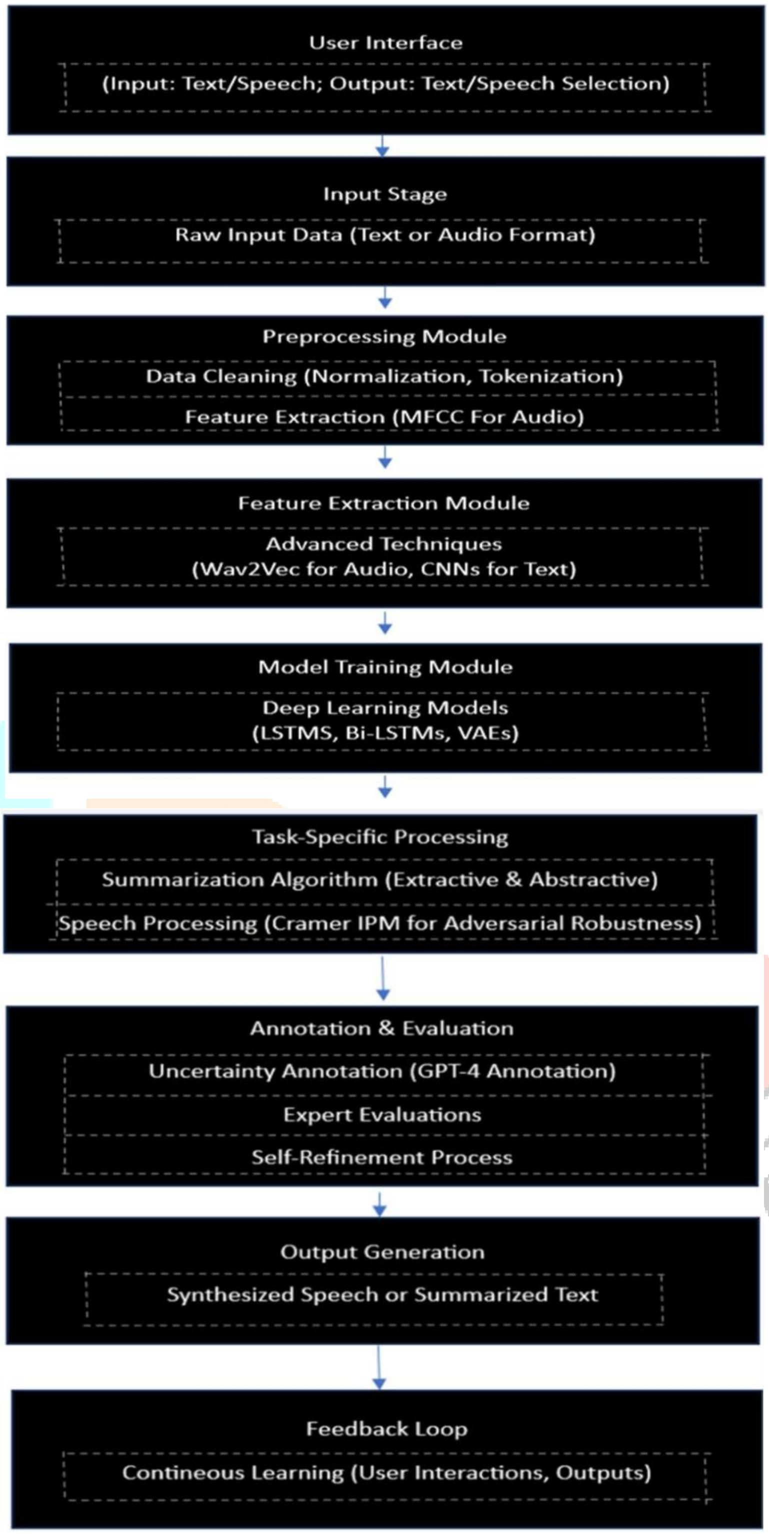
User Interface
(Input: Text/Speech; Output: Text/Speech Selection)
Input Stage
Raw Input Data (Text or Audio Format)
Preprocessing Module
Data Cleaning (Normalization, Tokenization)
Feature Extraction (MFCC For Audio)
Feature Extraction Module
Advanced Techniques
(Wav2Vec for Audio, CNNs for Text)
Model Training Module
Deep Learning Models
(LSTMS, Bi-LSTMs, VAEs)
Task-Specific Processing
Summarization Algorithm (Extractive & Abstractive)
Speech Processing (Cramer IPM for Adversarial Robustness)
Annotation & Evaluation
Uncertainty Annotation (GPT-4 Annotation)
Expert Evaluations
Self-Refinement Process
Output Generation
Synthesized Speech or Summarized Text
Feedback Loop
Contineous Learning (User Interactions, Outputs)

FIG 1 PROCESS FLOW OF MULTIMODAL TEXT AND SPEECH CONVERSION WITH SUMMARIZATION

Abbreviations and Acronyms

Text-to-speech, Speech-To-Text, Brain-Computer Interface, Cross-Lingual Translation, Summarization, Large Language Models, Deep Learning, Digital Signal Processing, Sequence-To-Sequence Models, Adversarial Algorithms, And Human-Computer Interaction.

2.1 List of Algorithms

Table 1 Algorithms

S.No	Algorithm	Description
1	CTX-txt2vec	Improves speech context understanding by creating semantic tokens from input text using VQ-diffusion in conjunction with Transformer blocks [8].
2	CTX-vec2wav	Converts predicted tokens to audio waveforms using contextual vocoding, improving speech flow and coherence [8].
3	VQ-diffusion	Applies a Markovian model to discrete speech data to optimize token prediction, essential for accurate, context-rich speech synthesis [8], [18].
4	Multi-task Warmup	Gradual training approach for vocoding models across tasks, aiding naturalness and consistent audio quality throughout training [2], [18].
5	Feature Extraction	Uses vq-wav2vec to extract semantic tokens, capturing critical speech characteristics while balancing articulation and acoustic fidelity [3].
6	Contextual Tokens	Integrates surrounding token data to enhance coherence in speech generation, beneficial for context-sensitive outputs [8].
7	Auxiliary Features	Includes pitch, energy, and probability of voice (POV), refining audio quality and making speech synthesis sound natural [2], [18].
8	Convolutional Neural Network (CNN)	Processes local dependencies in shorter text sequences, capturing word relationships within a limited context for effective feature extraction [9].
9	Long Short-Term Memory (LSTM)	Captures long-term dependencies in text sequences, making it suitable for contextual understanding across multiple sentences [5], [23].
10	Bidirectional LSTM (Bi-LSTM)	Reads sequences in both forward and backward directions, capturing context from both preceding and following words for improved understanding [23].
11	Gated Attention Graph Neural Network (GA-GNN)	Treats each word as a node in a graph structure to capture intricate semantic relationships, highlighting critical features for summaries [23].
12	Wav2Vec	Self-supervised ASR model that learns from unlabeled audio data, generating detailed representations that enhance speech-to-text transcription accuracy [3], [26].
13	Mel-Frequency-Cepstral Coefficients (MFCC)	Captures key speech characteristics from audio signals, making them suitable for SVM-based speech classification and recognition tasks [9].
14	Extractive Summarization	Selects essential sentences directly from the original text, creating a concise summary by focusing on key points without rephrasing [11], [5].
15	Abstractive Summarization	Generates a summary by rephrasing the main points, using models trained to create concise, coherent text that conveys essential information [11], [5].
16	Two-Tier Taxonomy	Categorizes uncertainty into lexical and semantic levels, helping to identify different expressions of uncertainty in both source texts and their summaries [7].

17	GPT-4 Annotation System	Utilizes GPT-4 to automatically annotate datasets with uncertainty tags, with expert involvement to refine the initial tags for improved reliability [7], [19].
18	Post-Hoc Self-Refinement	Uses expert feedback for iterative self-correction of model annotations, allowing the model to enhance its tagging accuracy over multiple rounds [7].
19	Cramér Integral Probability Metric (IPM)	Measures the difference between original and adversarial audio samples, generating adversarial signals that closely resemble authentic speech patterns to withstand noise [6].
20	White-Box Attack Framework	Develops adversarial signals that remain robust against over-the-air playback attacks, tested on systems like DeepSpeech and Kaldi [6].
21	EOT-Free Framework	Avoids time-consuming Expectation Over Transformation (EOT) operations by applying Cramér-IPM, which reduces distortion while enhancing robustness of adversarial attacks [6].
22	Variational Autoencoder (VAE)	Creates high-quality speech by encoding text into frame-level representations, significantly improving speech synthesis quality to approximate human recordings [18], [8].
23	Differentiable Durator	A duration modeling technique that aligns text with frame-level speech representations, reducing mismatches between training and inference [8].
24	Bidirectional Prior/Posterior Module	Matches prior and posterior distributions of text and speech using a flow model, enhancing speech synthesis quality through better alignment [18], [8].
25	Memory-Based VAE	Streamlines posterior complexity using attention mechanisms, focusing on high-quality reconstruction of waveforms without overloading computational resources [18], [8].

2.2 TECHNIQUES IN MULTIMODAL SYSTEMS

The field of multimodal systems has expanded to include a variety of methods for enhancing the performance of text translation and summarization. The development of these systems generally encompasses four main phases:

2.2.1 Data Preprocessing

Preprocessing is a crucial first step that ensures raw data is uniform and ready for model input. Techniques like Tokenization and Lemmatization, refine text by converting words into base forms, enhancing consistency. Noise reduction [10], which removes unnecessary or inconsistencies, optimizing data for better model learning and accuracy.

2.2.2 Model Training Approaches

Different types of models serve specific roles in translation and summarization tasks. Sequence-to-sequence (Seq2Seq) models with attention mechanisms [1] handle complex language dependencies and ensure the generated output remains cohesive. Transformers such as T5 AND BERT capture longer dependencies and context, making them particularly useful for summarization applications [11]. Recent advancements incorporate multi-task learning, where models are trained across multiple tasks (like ASR and summarization) to enhance adaptability and improve output quality [1], [8].

2.2.3 Evaluation Metrics

Various metrics are used to gauge the performance of models, including BLEU and ROUGE scores, which compare generated outputs against reference texts and are commonly used for translation and summarization evaluation [7], [11]. BERTScore, which assesses semantic accuracy by evaluating how well the model captures meaning in relation to reference texts. Domain-specific metrics such as MEDCON for medical terminology and Domain Vocabulary Overlap (DVO), measure a model's capability to handle specialized language [19].

2.2.4 Advanced Techniques

A range of advanced techniques further optimizes multimodal systems: Fine-tuning with domain-specific data [7], [19] and using in-context examples to improve model accuracy for specialized tasks. Temperature settings during text generation and adversarial training in voice output creation enhance the naturalness and reliability of synthesized responses [6]. Cross-modal and transfer learning allows models to apply learned features from one domain (eg., text), thus broadening their functional scope [3], [9].

2.3 TECHNOLOGIES IN MULTIMODEL SYSTEM

2.3.1 Natural Language Processing Tools

NLP tools are essential in text-based and multimodal systems, allowing machines to effectively analyze, interpret, and generate human language. These tools handle a range of linguistics tasks, including tokenization, stemming, named entity recognition (NER), sentiment analysis, and text summarization. By transforming unstructured language input into structured data, NLP enables downstream processes such as translation, summarization, speech synthesis, and more advanced interactions between humans and computers [5].

Modern frameworks like spaCy, NLTK, and Hugging Faces Transformers provide pre-trained models that streamline fundamental NLP operations. These libraries offer capabilities such as part-of-speech tagging, tokenization [5], and semantic embedding generation, which allows systems to extract meaningful insights from raw text and perform complex linguistics analysis. Tools like Gensim’s Word2Vec also create word embeddings that capture intricate semantic relationships, improving text comprehension and aiding in developing coherent summaries [10].

2.3.2 Speech Recognition Systems

Speech recognition systems play a crucial role in converting spoken language into text, making them integral to multimodal workflows, particularly in automatic speech recognition (ASR) [6]. By transcribing audio inputs, these systems enable smooth interactions and are widely used in applications such as virtual assistants, transcription services, and language learning platforms. Notable technologies include Google Speech-to-text and IBM Watson Speech-to-text, which provide highly accurate cloud-based transcription services. Additionally, open-source tools like Kaldi, DeepSpeech, and Lingvo are valued for their neural network-based architectures [6] and modular design, making them ideal for both research and customization. The HuBERT speech encoder further enhances detailed phonetic and audio features, improving accuracy even in challenging acoustic environments [3].

2.3.3 Machine Translation

These frameworks are essential for enabling smooth communication across different languages by converting text or speech from one language to another. Neural Machine Translation (NMT) systems, powered by deep learning, significantly improve translation quality by capturing subtle nuances and contextual meaning across languages [10]. Leading platforms like Google Translate and Microsoft Translate apply advanced NMT techniques to deliver more accurate translations [10]. Additionally, SpeechLM aligns speech and text by employing tokenizer to generate shared representations, enabling better integration of multimodal inputs [3]. This ensures seamless interaction across languages, even when systems involve both spoken and written components. MT frameworks are crucial in reducing language barriers, making real-time communication more accessible in glbal and multigoal settings. They also enhance user experience by enabling instant translation of sopoken content, supporting smoother interaction and accessibility for diverse users.

2.3.4 Text-to-speech (TTS) Synthesis

Text-to-speech (TTS) synthesis transforms text into natural-sounding speech, enhancing user engagement and accessibility in multimodal systems. It is widely applied in virtual assistants, screen readers, and customer support tools to provide auditory feedback. Advanced TTS models, like SPEAR-TTs, TextrolSpeech, and Salle, generate diverse audio outputs using techniques such as Residual Vector Quantization (RVQ) and self-supervised learning. These systems can replicate various voice styles with minimal speaker input and leverage models like GPT-3.5-Turbo for style-based prompts [2], [4], [19].

2.3.5 Large Languages Models (LLMS) and InContext Learning (ICL)

Large language models (LLMS) are highly proficient in executing complex NLP tasks, including summarization, translation, and domain adaptation. They can learn from context and generalize across various tasks with minimal training, making them versatile tools for multimodal systems. Models such as LLaMA-2, Vicuna, and Falcon are evaluated for their incorrect learning capabilities, which allow them to perform effectively across different domains without extensive fine-tuning [7], [16]. Fine-tunning methods like LoRA

(Low-Rank Adaptation) and LNA (LayerNorm & Attention) optimize these models for specific tasks by efficiently adjusting parameters, ensuring high performance while minimizing computational demands.

2.3.6 Evaluation Frameworks for Summarization Performance

Evaluation frameworks are crucial for measuring the quality and accuracy of generated text summaries. These frameworks assess the generated text against reference summaries, determining how well the outputs capture essential information.

G-eval, a framework based on GPT-4 scoring [19], evaluates the fluency and coherence of generated summaries. Additionally, metrics like Rouge and BERTScore are used to gauge the relevance and similarity between the generated outputs and reference texts [5], [11]. The AdaptEval framework assesses how well LLMs adjust to specialized vocabularies across different domains [7].

2.3.7 Attention Graph Neural Networks (GA-GNN)

Gated Attention Graph Neural Network (GA-NNs) are designed to capture complex word relationships by modeling words as nodes in a graph [23] and connecting them through edges based on their associations. Built on the principles of Graph Neural Networks, GA-GNNs are adept at processing complex, non-linear data structures, making them ideal for tasks requiring deep text analysis. By combining global and local semantic information, GA-GNNs enable the system to grasp and interpret intricate relationships within text, making it particularly useful for tasks like summarization, sentiment analysis, and entity recognition. In multimodal systems, GA-GNNs improve understanding of context and meaning.

2.3.8 Adversarial Attack Algorithms for Speech Systems

Adversarial attack algorithms are utilized to evaluate the resilience of speech-to-text systems by generating deceptive audio inputs that closely resemble genuine signals, yet lead to transcription inaccuracies [6], [26]. These algorithms are essential for identifying vulnerabilities in automatic speech recognition (ASR) systems, ensuring they can resist intentional disruptions or background noise encountered in real-world applications.

Frameworks like DeepSpeech, Kaldi, and Lingvo are frequently tested against these adversarial inputs to assess their robustness [1]. Given that these systems are built on advanced neural architectures for accurate transcription, it is vital to understand their performance in adverse conditions. Conduction adversarial testing is crucial for guaranteeing that speech recognition systems remain reliable and secure.

2.3.9 Encoder-Decoder Models for Summarization

Encoder-Decoder models are instrumental in automatic text summarization, as they effectively capture the contextual relationships among words in input sequences. These models convert lengthy texts into concise summaries while preserving the original meaning and key information. Technologies like BART and PEGASUS-X excel in text summarization [5], [11] by leveraging attention mechanisms to grasp complex interrelationships among words and phrases. Furthermore, Bidirectional LSTM models improve the summarization process by considering both past and future contexts [23], ensuring that the generated summaries are coherent and meaningful.

2.3.10 End-to-End (E2E) Speech-to-Text Translation (S2TT)

End-to-end (E2E) models integrate automatic speech recognition (ASR) and machine translation (MT) into a single process [1], [16], reducing latency and minimizing the risk of error propagating through the workflow. By streamlining the translation process and eliminating intermediate steps, E2E models offer greater efficiency for real-time applications. The W2v-BERT speech encoder captures rich phonetic and speech features, facilitating effective integration between ASR and MT within the E2E framework [1]. This methodology enhances the accuracy of speech-to-text translations and ensures quicker responses in applications such as live translation and virtual meetings. E2E models significantly improve multimodal systems by providing seamless speech-to-text translation, making them particularly well-suited for real-time communication in multilingual environments.

2.4 DATASETS

2.4.1 Multilingual Text and Speech Translation Datasets

a. Common Crawl & Wikipedia: These large-scale corpora contain diverse multilingual data for training generalized translation and summarization models, covering various domains [13].

b. Europarl Corpus: Originating from European Parliament proceedings, this dataset provides parallel, structured text, suitable for developing consistent and formal translation systems across multiple languages [13].

c. IWSLT TED Talks Corpus: Multilingual TED talk transcriptions aid in real-time translation of conversational language, enhancing models intended for public speaking contexts [13].

2.4.2 Speech Datasets for ASR and TTS

a. LibriSpeech: A large English speech corpus with transcriptions, useful for ASR and TTS tasks with speaker adaptation and noise handling capabilities [9].

b. Mozilla Common Voice: A community-contributed dataset covering diverse accents, ages, and dialects, ideal for building speech recognition systems with high generalizability [9].

c. VoxPopuli: European Parliament speeches with multilingual annotations support multilingual summarization and translation, focusing on formal, spoken contexts [13].

2.4.3 Translation Datasets for Colloquial and Formal Contexts

a. OpenSubtitles: Subtitled movie and TV data provide conversational and colloquial language, including slang and regional dialects, making it valuable for translating informal text [13].

b. Tatoeba: A user-contributed multilingual corpus with sentences translated across many languages, supporting translation model generalization [13].

2.4.4 Summarization Datasets for News, Scientific, and Medical Domains

a. CNN/Daily Mail: A collection of news articles and summaries useful for extractive and abstractive summarization, enabling models to understand and summarize real-world news content [11].

b. XSum and Gigaword: Focusing on concise single-sentence summaries, these datasets aid in developing models that summarize text into brief, informative statements [11].

c. Scientific Summarization - Scisumm and ArXiv: These provide summaries of scientific articles and papers, ideal for models summarizing technical or research content [19].

d. MIMIC-III and MIMIC-CXR: De-identified clinical notes and radiology reports support summarizing patient data and generating medical reports, critical for healthcare applications [19].

e. MeQSum: This dataset includes health-related queries for patient questions, useful for simplifying and summarizing complex medical language for general understanding [19].

2.4.5 Cognitive and Interactional Datasets

a. ZuCo (Zurich Cognitive Dataset): Includes EEG and eye-tracking data, capturing brain activity during reading tasks. This dataset enables research on multimodal interpretation and brain-computer interfaces (BCIs) [17].

b. COCO (Common Objects in Context): A multimodal dataset with images paired with textual descriptions, supporting cross-modal translation tasks, where images or other sensory data accompany text [13].

III. EVALUATION METRICS

To thoroughly assess how well the model performs, a diverse set of quantitative and qualitative metrics was carefully chosen. These metrics provide insights into different qualities of the model's output, including how natural it sounds, how closely it resembles human speech, its clarity, and its overall effectiveness. Each metric is tailored to capture a unique aspect of the model's capabilities, whether in generating audio or text. Below is an overview of each metric, highlighting its purpose, how it's calculated, and why it's significant in evaluating the model's performance.

1. Mean Opinion Score (MOS)

Description: MOS assesses the perceived quality of generated speech by having listeners rate the naturalness and speaker similarity [2] on a 1-5 scale, with higher scores indicating closer resemblance to natural human speech.

Formula: $\mathrm{MOS} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Ri}$

2. Speaker Encoder Cosine Similarity (SECS)

Description: SECS quantifies the degree of similarity between the speaker characteristics in the generated and reference audio by comparing their embeddings [18]. Higher values indicate better alignment.

Formula: $\mathrm{SECS} = \frac{\mathrm{v1.v2}}{||\mathrm{v1}||||\mathrm{v2}||}$

3. Word Error Rate (WER)

Description: WER is used to evaluate transcription accuracy by calculating the number of errors in speech-to-text conversions [6]. A lower WER indicates fewer errors and more accurate transcriptions.

Formula: $\mathrm{WER} = \frac{\mathrm{S+D+I}}{\mathrm{N}}$

4. Comparative Mean Opinion Score (CMOS)

Description: CMOS directly compares the generated speech quality against human recordings [19] on a 7-point scale. Scores near zero suggest minimal perceptual differences between generated and human speech.

5. Wilcoxon Signed-Rank Test

Description: This statistical test assesses whether there is a significant difference in quality between generated speech and human recordings [9]. When used alongside CMOS, it helps confirm quality equivalence.

Formula: $\mathbf{W} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{sign(di).\,Ri}$

6. Soft Dynamic Time Warping (Soft-DTW) Loss

Description: Soft-DTW measures alignment accuracy by comparing frame-level differences during training. This is used to optimize text-speech alignment, with lower scores indicating better alignment.

Formula: $\mathrm{soft - DTW(X, Y)} = \sum_{\mathrm{ij}} \exp(-\frac{|\mathrm{xi-yj}|}{\sigma})$

7. ROUGE Score

Description: ROUGE evaluates the content similarity of generated summaries to reference summaries using n-gram overlap, commonly applied in text summarization tasks [11].

Formula: $\mathrm{ROUGE - N} = \frac{\sum_{\mathrm{gram \in generated}} \mathrm{count(gram) \cap count(reference)}}{\sum_{\mathrm{gram \in reference}} \mathrm{count(gram)}}$

8. BERT Score

Description: BERTScore assesses contextual similarity between generated and reference text by calculating the similarity of embeddings for each token [11], which provides insights into fluency and coherence.

Formula: $\mathrm{BERTScore(x, y)} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{maxcosine\_similarity(e(xi), e(yi))}$

9. GPT-4-based G-eval

Description: This metric leverages GPT-4 to provide qualitative evaluations, focusing on aspects like coherence [7], [19], fluency, and domain alignment, giving a comprehensive overview of generated text quality.

10. Manual Evaluation with Cohen’s Kappa

Description: Expert evaluators assess the quality of domain adaptation. Cohen’s kappa (κ\kappaκ) measures agreement among raters, with higher values indicating greater consensus [7], [19].

Formula: Pe =i=1∑k (pi,1 · pi,2)

11. Sentence-Level Accuracy

Description: This metric tracks how accurately sentences are transcribed following adversarial perturbations. Lower accuracy indicates the strength of the perturbations in disrupting the transcription [6].

Formula: $\mathrm{Accuracy} = \frac{\text{Number of correct sentences}}{\text{Total number of sentences}}$

12. Robustness Across Playback Conditions

Description: This test evaluates resilience of generated audio under varied playback conditions, such as noise and distortion, and assesses how well the model handles different real-world scenarios [6], [3].

$$\textbf{Formula: } \text{Robustness} = \frac{\text{Accuracy under noise conditions}}{\text{Accuracy under clean conditions}}$$

13. Cramér Integral Probability Metric (Cramér-IPM)

Description: Cramér-IPM measures the similarity between original and adversarial signal distributions, ensuring that adversarial modifications remain close to the natural speech space [6].

Formula: Crame´r-IPM(P,Q)=f∈Fsup |Ex∼P [f(x)]−Ex∼Q [f(x)]|

14. Accuracy of Annotation Elements

Description: Lexical and semantic tagging accuracy, such as part-of-speech tagging and uncertainty labeling, is assessed by experts to ensure that annotations reflect accurate linguistic elements [7].

$$\textbf{Formula}: \text{Accuracy} = \frac{\text{Correct annotations}}{\text{Total annotations}}$$

15. Model Self-Refinement Accuracy

Description: This metric tracks the model's improvement in accuracy after each round of self-refinement. With each iteration, accuracy should increase, demonstrating the effectiveness of refinement [7].

$$\textbf{Formula}: \text{Accuracy\_refined} = \frac{\text{Correct predictions after refinement}}{\text{Total predictions after refinement}}$$

16. Uncertainty Transfer Fidelity

Description: This assesses the retention of uncertainty expressions from original text to generated summaries, focusing on how well the model maintains these nuances in the output [7].

$$\textbf{Formula}: \text{Fidelity} = \frac{\text{Uncertainty preserved in summary}}{\text{Uncertainty in original text}}$$

IV. Challenges & Limitations

Multimodal systems for translation and summarization bring a range of unique challenges that are essential to address for their effectiveness and scalability. One major limitation lies in the difficulty of retaining context, as these models often struggle to capture cultural nuances, idiomatic expressions, and specialized terminology [12]. This can lead to misinterpretations or overly simplified outputs that miss the subtleties of meaning, especially when translating across languages with varied structures and expression styles.

Another significant challenge is the availability of comprehensive data resources, especially for less commonly spoken languages. Many languages lack extensive digital resources, limiting representation for certain dialects and vocabularies in multilingual models. This can introduce biases and affect the quality of translations, particularly for languages classified as low-resource. Integrating text, audio, and visual data further complicates things, as achieving synchronization across modalities is complex. It requires precise alignment to ensure, for instance, that spoken text in one language matches visual data or subtitling in another [17]. Synchronizing such data accurately becomes particularly challenging in noisy environments or when processing diverse accents.

Additionally, multimodal systems are highly resource-intensive, requiring considerable computational power to process data in real-time [1], [5]. This is especially true when these models are fine-tuned to accommodate specific domains or vocabularies, which adds to the hardware and software complexity. Another difficulty in these systems is error propagation, where errors in one stage such as inaccuracies in automatic speech recognition (ASR) can impact subsequent processing steps, ultimately affecting the final quality of translations or summaries [6].

Evaluating the quality of outputs from multimodal systems also presents challenges. Measuring aspects like translation accuracy, semantic coherence, and audio quality often depends on subjective human judgments, which can vary considerably. Establishing objective evaluation standards that accurately capture the quality of multimodal content is complex but essential. Addressing these issues is critical to making multimodal systems more reliable and scalable for a wide range of applications.

V. Result Discussion

The body of research demonstrates notable advancements in multilingual text translation, speech-to-text, and summarization, shedding light on both significant achievements and ongoing challenges. Transformer models, such as mBERT and XLM, have proven effective in addressing low-resource languages, facilitating translations where data is scarce [11], [13]. However, fine-tuning these models for languages with abundant resources poses challenges, particularly the risk of overfitting, which can restrict their adaptability to different linguistic frameworks [1].

In the domain of speech-to-text (S2T) technology, tools like Wav2Vec have significantly improved transcription accuracy, especially in challenging, noisy environments. Utilizing unsupervised learning, these models adeptly manage variations in audio quality, but they also demand considerable computational power, which may limit their use in settings with fewer resources. Summarization research shows that various methods cater to different requirements: extractive summarization excels in processing longer, unstructured texts by extracting key sentences directly [11], thus retaining the original meaning. In contrast, abstractive summarization transforms the information into shorter, more natural summaries, making it suitable for concise texts. However, while abstractive methods offer greater flexibility, they may also lead to minor inaccuracies or misinterpretations [1], [11].

Cross-lingual summarization stands out as a particularly valuable approach, as integrating translation and summarization into a unified process often results in more coherent and fluid summaries compared to treating the tasks separately [12]. Still, these systems must be careful to account for cultural differences and linguistic nuances to prevent miscommunication, given that subtle meanings in one language might not translate seamlessly into another.

Furthermore, the use of uncertainty annotations introduces an essential layer of transparency, indicating the confidence level of the model's outputs. This feature is especially useful in multilingual and cross-domain applications, helping users gauge the reliability of the system and identify when human input might be necessary. Overall, the literature underscores the need to balance accuracy with computational efficiency, aiming to develop models that are both effective and user-friendly across a wide range of languages and contexts. This dynamic field looks forward to a future enriched by multimodal and multilingual applications, emphasizing accessibility and inclusivity [7].

VI. Conclusion

In conclusion, the advancement of multimodal systems presents numerous opportunities for enhancing their efficiency, inclusivity, and security. Developing noise-resistant models will enable these systems to perform reliably in challenging auditory environments, while targeted efforts to support low-resource languages through techniques like data augmentation and cross-lingual learning will help bridge linguistic gaps.

Additionally, incorporating contextual embeddings can refine translation accuracy by adapting to user interactions, and improving explainability will foster greater trust, particularly in critical fields such as healthcare and law. Strengthening defenses against adversarial attacks in speech recognition and refining adaptive neural architectures for speech synthesis will further enhance reliability and performance across diverse applications. As research in these areas progresses, multimodal systems will continue to evolve, offering more accessible, robust, and practical solutions for real-world challenges.

VII. References

[1]Challagundla, B. C., & Peddavenkatagari, C. R. (2024). Neural Sequence-to-Sequence Modeling with Attention by Leveraging Deep Learning Architectures for Enhanced Contextual Understanding in Abstractive Text Summarization. International Journal of Computational Intelligence and Applications.

[2]Tan, X., Chen, J., Liu, H., & Cong, J. (2024). NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[3]Zhang, Z., Chen, S., Zhou, L., Wu, Y., Ren, S., Liu, S., Yao, Z., Gong, X., Dai, L., Li, J., & Wei, F. (2024). SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[4]Ji, S., Zuo, J., Fang, M., Jiang, Z., Chen, F., Duan, X., Huai, B., & Zhao, Z. (2024). TEXTROLSPEECH: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[5]Ji, S., Zuo, J., Fang, M., Jiang, Z., Chen, F., Duan, X., Huai, B., & Zhao, Z. (2024). TEXTROLSPEECH: A Text Style Control Speech Corpus with Codec Language Text-to-Speech Models. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[6]Esmaeilpour, M., Cardinal, P., & Koerich, A. L. (2024). Towards Robust Speech-to-Text Adversarial Attack. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[7]Huang, C.-W., Lu, H., Gong, H., Inaguma, H., Kulikov, I., Mavlyutov, R., & Popuri, S. (2024). Investigating Decoder-Only Large Language Models for Speech-to-Text Translation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[8]Afzal, A. (2024). AdaptEval: Evaluating Large Language Models on Domain Adaptation for Text Summarization. Technical Report, Technical University of Munich.

[9]Du, C., Guo, Y., Shen, F., Liu, Z., Liang, Z., Chen, X., Wang, S., Zhang, H., & Yu, K. (2024). UniCATS: A Unified Context-Aware Text-to-Speech Framework with Contextual VQ-Diffusion and Vocoding. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[10]Rabiner, L. R., & Schafer, R. W. (2024). Introduction to Digital Speech Processing. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[11]Ilango, B. (2024). A Machine Translation Model for Abstractive Text Summarization Based on Natural Language Processing. International Journal of Computational Linguistics.

[12]Zhu, J., Zhou, Y., Zhang, J., & Zong, C. (2024). Attend, Translate and Summarize: An Efficient Method for Neural Cross-Lingual Summarization. Proceedings of the Conference on Neural Information Processing Systems (NeurIPS).

[13]O'Mahony, J., Lai, C., & King, S. (2022). Combining Conversational Speech with Read Speech to Improve Prosody in Text-to-Speech Synthesis. In H. Ko & J.H.L. Hansen (Eds.), Proceedings of Interspeech 2022. Interspeech - Annual Conference of the International Speech Communication Association (ISCA), 3388-3392.

[14]Alaudinova, D. (2024). Written Translation of Texts Related to Different Spheres. Journal of Translation Studies.

[15]Mingote, V., Gimeno, P., Vicente, L., Khurana, S., Laurent, A., & Duret, J. (2024). Direct Text-to-Speech Translation System Using Acoustic Units. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[16]Saeki, T., Maiti, S., Li, X., Watanabe, S., Takamichi, S., & Saruwatari, H. (2024). Learning to Speak from Text: Zero-Shot Multilingual Text-to-Speech with Unsupervised Text Pretraining. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[17]Le, C., Zhou, L., & Zeng, M. (2024). ComSL: A Composite Speech-Language Model for End-to-End Speech-to-Text Translation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[18]Duan, Y., Zhou, J., Wang, Z., Wang, Y.-K., & Lin, C.-T. (2024). DeWave: Discrete EEG Waves Encoding for Brain Dynamics to Text Translation. Proceedings of the International Conference on Brain-Computer Interface and Applications.

[19]Li, Y. A., Vinay, S. R., Mischler, G., & Mesgarani, N. (2024). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[20]Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. S. (2024). Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. Journal of Medical Informatics, 43(3), 456-465.

[21]Kharitonov, E., Girgin, S., Vincent, D., Pietquin, O., Borsos, Z., Marinier, R., Sharifi, M., Tagliasacchi, M., & Zeghidour, N. (2024). Speak, Read, and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision—proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[22]Vinnarasu, A., & Jose, D. V. (2024). Speech-to-Text Conversion and Summarization for Effective Understanding and Documentation. Proceedings of the International Conference on Computational Intelligence in Communication Networks (CICN).

[23]Wang, M., Shafran, I., Soltau, H., Han, W., Cao, Y., Yu, D., El Shafey, L. (2024). Speech-to-Text Adapter and Speech-to-Entity Retriever Augmented LLMs for Speech Understanding. Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech).

[24]Huang, J., Wu, W., Li, J., & Wang, S. (2024). Text Summarization Method Based on Gated Attention Graph Neural Network. Proceedings of the International Conference on Natural Language Processing (ICON).

[25]Monteiro, R., & Pernes, D. (2024). Towards End-to-End Speech-to-Text Summarization. Proceedings of the European Conference on Artificial Intelligence (ECAI).

[26]Korchynskyi, V. V., & Vynogradov, I. V. (2024). Methods of Improving the Quality of Speech-to-Text Conversion. Proceedings of the International Conference on Speech Technology and Applications (SPECOM).

IJCRT