Text-to-Speech Generation with Multilingual Assistance and AI Integration for Information Generation

Dr. JayaSudha K[1*] Appu Gowda B S[2], Meghana B R[3], Geetha U[4], Ashritha G N[5]

[1*] Prof & Head, Department of AIML, Sri Krishna Institute of Technology, Bengaluru, INDIA

[2,3,4,5] UG Students, Sri Krishna Institute of Technology, Bengaluru – INDIA

Abstract—Multilingual speech technologies are critical to enabling seamless communication across languages, empowering users with diverse linguistic backgrounds, and supporting accessibility in digital systems. This paper presents a complete multilingual Text-to-Speech (TTS) architecture enhanced with Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and AI-driven information extraction. Unlike conventional TTS systems that simply convert text into speech, this integrated pipeline analyzes speech, extracts meaningful information, translates content to a target language, and finally generates expressive and natural audio. We describe each subsystem in detail, the engineering decisions behind the composite pipeline, and the challenges encountered in multilingual deployment settings.

Keywords- Automatic Speech Recognition (ASR), Hidden Markov Model-Gaussian Mixture Model (HMM-GMM), Neural Machine Translation (NMT), Natural Language Processing (NLP), Open Platform for Unified Speech (OPUS), Representational State Transfer (REST), Text to Speech (TTS)

1.INTRODUCTION

Speech is one of the most natural modes of communication, and modern computing systems increasingly rely on speech interfaces for accessibility, automation, and interaction. With the rise of multilingual users on global platforms, the need for speech technologies that can understand and generate content across languages has grown substantially. Traditional TTS systems served monolingual purposes and often lacked expressive flexibility. Recent deep learning approaches have significantly improved the naturalness, rhythm, and clarity of synthesized speech.

A multilingual TTS system, however, requires more than converting text into speech. It must integrate speech recognition [1], translation, and semantic extraction, especially when Transforming speech from one language into another. The process involves several challenges: handling accents, code-mixed speech, domain-specific vocabulary, and contextual emphasis. The proposed work addresses these issues using a modular pipeline that can be adapted or scaled based on performance requirements and available computational resources.

The objective of this paper is to present a well-structured architecture that combines state-of-the-art ASR, NMT, and TTS components. Each module contributes uniquely to the overall system and has been engineered to ensure high-quality output speech, even in multilingual and noisy environments. By following a systematic IEEE-style document structure, this paper delivers technical clarity, academic rigor, and engineering depth appropriate for student and researcher audiences.

2.RELATED WORK

Multilingual speech technologies have gained considerable attention as digital communication increasingly demands seamless interaction across languages [2]. Systems that combine Automatic Speech Recognition (ASR), Neural Machine Translation (NMT), and Text-to-Speech (TTS) enable speech-to-speech communication and form the backbone of multilingual voice-based applications. Research in each of these domains has progressed independently, but recent efforts emphasize their integration for real-world deployment.

Automatic Speech Recognition- Traditional ASR systems were constructed using Hidden Markov Models (HMMs) in combination with Gaussian Mixture Models (GMMs) [3]. These approaches relied heavily on manually designed acoustic features and required careful tuning for each language and speaker group. As a result, they struggled to generalize across accents, noisy environments, and diverse linguistic patterns, particularly in multilingual scenarios.

The introduction of deep learning techniques marked a major improvement in speech recognition [4] performance. Deep neural networks enabled automatic feature learning, reducing the dependence on handcrafted representations. More recently, end-to-end ASR models have become prominent by directly mapping speech signals to textual output. Self-supervised learning approaches, which exploit large amounts of unlabeled audio data, further enhanced recognition accuracy and robustness [5] [6]. Models trained in this manner demonstrated strong adaptability to new languages and low-resource settings, making them suitable for multilingual speech applications [7] [8]. Large-scale multilingual training has also improved generalization across varying acoustic and linguistic conditions [9].

Text-to-Speech Synthesis- Early TTS systems employed rule-based synthesis or statistical parametric techniques, which often resulted in unnatural and monotonous speech. With the advent of neural networks, text-to- speech synthesis evolved into an end-to-end learning problem. Sequence-to-sequence models enabled direct prediction of acoustic features from text, significantly enhancing speech naturalness and intelligibility [10].

Subsequent improvements focused on reducing inference time and increasing stability during synthesis. Non- autoregressive models introduced explicit control over speech duration and prosody, leading to faster and more consistent output. Recent end-to-end architectures integrate acoustic modeling and waveform generation into a single framework, achieving highly natural speech quality with lower latency. Multilingual TTS systems extend these ideas by learning shared phonetic and acoustic patterns across languages, enabling effective speech generation for multiple linguistic contexts [11].

Integrated Multilingual Speech Pipelines- Although ASR, NMT, and TTS have individually achieved high performance, their integration into a unified multilingual pipeline remains an evolving research area [12]. Conventional pipeline-based systems often suffer from cumulative errors and increased latency as output from one module becomes input to the next [13]. Recent studies explore tighter coupling between components to improve efficiency and reduce information loss. Despite these advances, challenges persist in achieving real- time performance, handling language switching, and supporting low-resource languages effectively [14]. Many existing systems are evaluated in controlled or offline environments, limiting their practicality for real-world multilingual communication [15].

Neural Machine Translation- Machine translation systems initially followed rule-driven and statistical paradigms, both of which faced limitations in handling complex linguistic structures and contextual dependencies. The shift toward neural machine translation introduced encoder–decoder architectures that learned translation patterns directly from data, leading to more fluent and context-aware outputs [16].

The emergence of attention-based mechanisms significantly improved translation quality by allowing models to focus on relevant parts of the input sequence. Transformer-based architectures further advanced this field by enabling parallel processing and efficient modeling of long-distance relationships within text. In multilingual settings, a single translation model is often trained on multiple language pairs using shared representations. This strategy promotes knowledge transfer across languages, reducing the need for extensive parallel datasets and improving translation performance for underrepresented languages. Such multilingual NMT systems are particularly effective when integrated into speech-based pipelines [17].

3. SYSTEM ARCHITECTURE AND MECHANISM

The system architecture consists of three main modules: ASR, NMT, and TTS. Audio input is first processed by the ASR module to generate accurate transcriptions across multiple languages and accents. The transcribed text is then translated by the NMT module, ensuring contextual correctness and cultural relevance. Finally, the TTS module synthesizes natural, expressive speech from the translated text, incorporating prosody and emotion modeling for enhanced user experience.

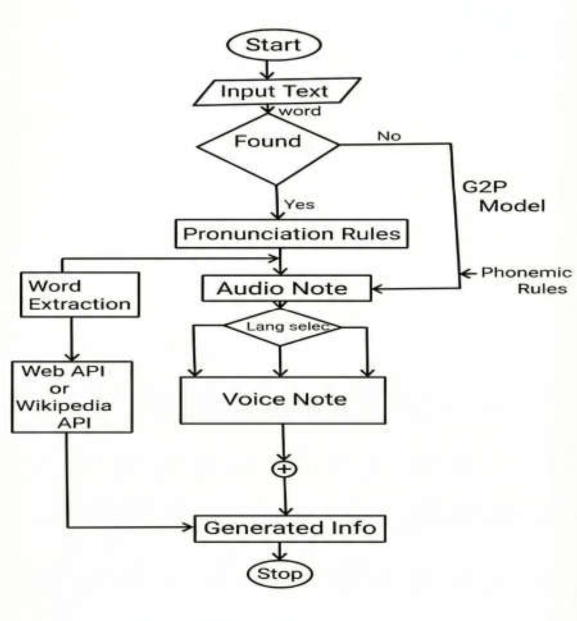


Fig. 1: Flowchart of Text to Speech model

The given diagram shown in Fig.1 represents the workflow of a Multilingual AI Text-to-Speech (TTS) system integrated with AI-based information retrieval. The process begins when the user provides input text or a query, which may contain general words, technical terms, or domain-specific phrases. The system first checks whether the input word or phrase exists in the predefined linguistic corpus. If it is available, the system directly proceeds with learned phonetic representations; otherwise, standard pronunciation rules are applied to ensure correct phoneme generation for unseen or rare words.

Once the text is validated, it is passed to an AI Grapheme- to-Phoneme (G2P) model, which converts written text into phonemic sequences. This conversion is crucial for handling multilingual inputs, as pronunciation rules vary significantly across languages. The system applies both standard phonemic rules and contextual phonemic rules, allowing it to adjust pronunciation based on sentence context, word position or surrounding phonemes. This step improves naturalness and intelligibility, especially in languages with complex pronunciation patterns.

The processed phonemes are then forwarded to the AI Voice Synthesis module, where the user-selected language is used to generate speech. This module supports multiple AI voices and accents, enabling regional accent adaptation and speaker diversity. Simultaneously, the system integrates an AI Information Retrieval module, which fetches relevant contextual information from knowledge graphs or web APIs when the input query requires factual or explanatory content. The retrieved information is summarized and synchronized with the speech output.

Finally, the system produces a combined output consisting of synthesized audio along with an AI-generated information summary, delivering both spoken content and enriched knowledge to the user. This integrated approach makes the system suitable for applications such as virtual assistants, educational tools, multilingual help systems, and accessibility platforms, ensuring accurate pronunciation, contextual awareness, and informative responses across multiple languages.

User Context Modelling:

The use case diagram shown in Fig.2 represents the functional workflow of an AI- assisted Text-to-Speech (TTS) system, illustrating how user input is processed, enhanced using artificial intelligence, and finally converted into speech output.

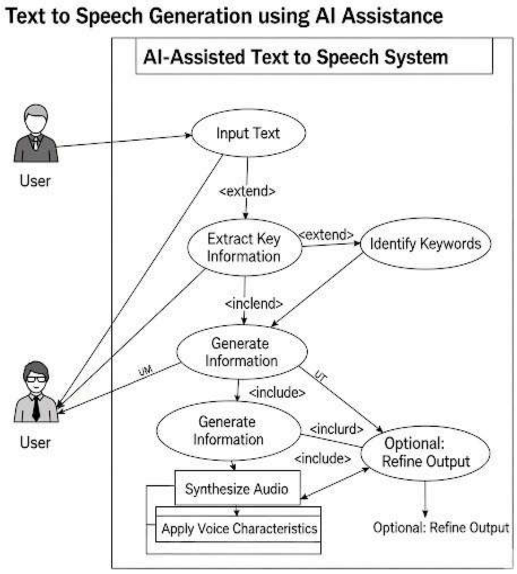


Fig. 2: Use case diagram of Text to Speech model.

1.User Input

The process begins when the user provides textual input to the system. This input can be plain text, queries, or informational content that the user wants to convert into speech. The system acts as an intelligent interface between the user and the speech synthesis engine.

2.Input Text Processing

Once the text is received, the system initiates input analysis. This step ensures that the text is structured properly and ready for further processing. The system may perform basic preprocessing such as cleaning, tokenization, or sentence segmentation.

3.Extraction of Key Information

The AI module then extracts key information from the input text. This step focuses on identifying important entities, topics, or meaningful segments within the content. This allows the system to understand the context and intent of the text rather than processing it as raw data.

4.Keyword Identification

As an extension of key information extraction, the system identifies keywords that play a crucial role in determining emphasis, tone, or relevance. These keywords help guide the AI in generating accurate and context-aware speech output.

5.AI-Based Information Generation

Based on the extracted information and keywords, the system performs AI-driven information generation. This stage may include expanding short inputs, generating additional explanations, or restructuring content to improve clarity and coherence before speech synthesis.

6.Optional Output Refinement

The system includes an optional refinement stage, where the generated information can be fine-tuned. This may involve adjusting sentence structure, removing redundancy, or improving linguistic quality to produce more natural and fluent speech output.

7.Speech Synthesis

After finalizing the text content, the system proceeds to synthesize audio. The refined text is converted into speech using a Text-to-Speech engine. This stage focuses on generating intelligible and smooth audio output.

8.Application of Voice Characteristics

To enhance user experience, voice characteristics such as pitch, speed, tone, and accent are applied. This allows customization of the synthesized speech, making it more natural and suitable for different user preferences or application requirements.

9.Final Audio Output

The system delivers the final synthesized speech to the user. The output reflects both the original input intent and the AI- assisted enhancements applied during processing.

The architecture begins with audio pre-processing, where raw speech is denoised and segmented. The ASR module converts speech to text while simultaneously identifying the spoken language. An information extraction engine identifies key terms, entities, and structural cues in the text. These extracted insights help shape translation quality by ensuring that entity names, technical terms, or context-dependent expressions are not mistranslated.

The translated text is then passed into the TTS module, which generates an expressive and natural-sounding waveform. Prosody embeddings guide pitch, rhythm, and speaking style. The system supports multiple voice models and emotional conditions, enabling dynamic speech generation tailored to user needs.

Real-Time User context Engine:

The sequence diagram shown in Fig.3 represents the flow of TTS model with the following Key Components:

1.User: The person initiating the request.

2.App Interface: The frontend (mobile or web) that handles the user interaction.

3.Multilingual AI Service: A Large Language Model (LLM) or AI engine (like Gemini) that understands and generates text in various languages.

4.TTS Engine: The Text-to- Speech service that converts written text into natural-sounding audio.

5.Audio Processing: The TTS Engine processes the text and returns a playable audio file.

6.Output: The app plays the audio response back to the user.

4.DATA PREPROCESSING

Constructing effective multilingual speech systems requires meticulous dataset collection, preprocessing, and augmentation to handle the immense variability in languages, accents, and environmental conditions. Automatic Speech Recognition (ASR) models depend on high-quality speech corpora such as Common Voice and LibriSpeech, which provide coverage for multiple accents, speaking styles, and real-world recording conditions, while parallel text datasets like OPUS and IndicNLP enable Neural Machine Translation (NMT) to handle multilingual translation tasks. Text-to-Speech (TTS) models, on the other hand, utilize datasets such as LJSpeech, LibriTTS, or carefully curated Indic recordings to capture phonetic richness, speaker diversity, and natural prosody. Preprocessing forms the backbone of these systems, including audio normalization, noise reduction, segmentation, and precise alignment of speech-text pairs. Consistent text representation across languages is achieved using subword tokenization methods like SentencePiece or Byte-Pair Encoding (BPE), which allow models to handle out- of-vocabulary words and morphologically rich languages. For low-resource languages, data augmentation techniques such as pitch and speed perturbations, background noise injection, and synthetic sentence generation through back-translation significantly enhance model robustness and generalization.

Beyond preprocessing, advanced refinements ensure that multilingual systems are resilient and performant in real-world scenarios. Accent-aware dataset balancing prevents models from being biased toward dominant dialects, while energy normalization ensures consistent loudness across recordings, which is critical for TTS naturalness. Phoneme-level alignment, particularly for languages without standardized orthography, guarantees accurate pronunciation and smoother prosody. Cross- lingual transfer learning allows models to leverage knowledge from high-resource languages, boosting performance for low- resource languages with limited training data. Additionally, engineering considerations such as scalable data pipelines, batch processing, and efficient GPU utilization enable the training of models across dozens of languages simultaneously. Evaluating systems under diverse conditions, including code- switching, noisy environments, and varied speaker profiles, ensures robustness and reliability. Together, these strategies create a solid foundation for multilingual ASR, NMT, and TTS systems, delivering accurate recognition, natural speech synthesis, and consistent translation performance across languages, dialects, and real-world scenarios.

5.IMPLEMENTATION DETAILS

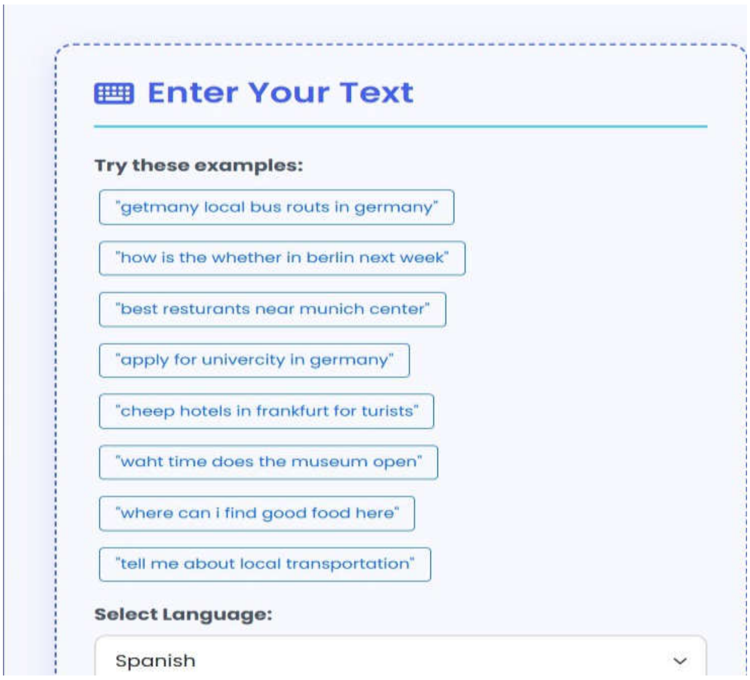
The proposed multilingual TTS system is implemented using a modular microservice architecture, enabling independent scaling, debugging, and optimization of each subsystem. The ASR, information extraction, NMT, and TTS components run as individual services communicating through REST APIs or lightweight gRPC calls. This ensures that performance bottlenecks in one stage do not affect the entire pipeline, and upgrades to one model can be seamlessly integrated without reconfiguring the entire system.

From an engineering standpoint, the implementation prioritizes real-time performance and low latency. ASR inference uses mixed precision and ONNX acceleration, reducing compute overhead. Whisper-based ASR is optimized through audio chunking, avoiding delays inherent in processing long audio streams. Further, model quantization is applied to compress weights with minimal accuracy loss, enabling deployment on devices with limited hardware resources. These optimizations result in smoother streaming and faster end-to-end response times, essential for user-facing applications.

For translation, transformer models are deployed with beam search decoding to enhance consistency in longer sentences. Domain-adaptive fine-tuning ensures that the model remains sensitive to proper nouns and technical terminology. The TTS implementation leverages FastSpeech2 for fast inference but uses VITS for high-quality synthesis when latency requirements are moderate. To support emotion and style variations, prosody embeddings are trained on curated expressive datasets. This enables dynamic control of speaking speed, pitch, and emotional tone. The entire pipeline is orchestrated using Docker containers, allowing cloud or on-premise deployment depending on security requirements.

6. EVALUATION AND RESULTS

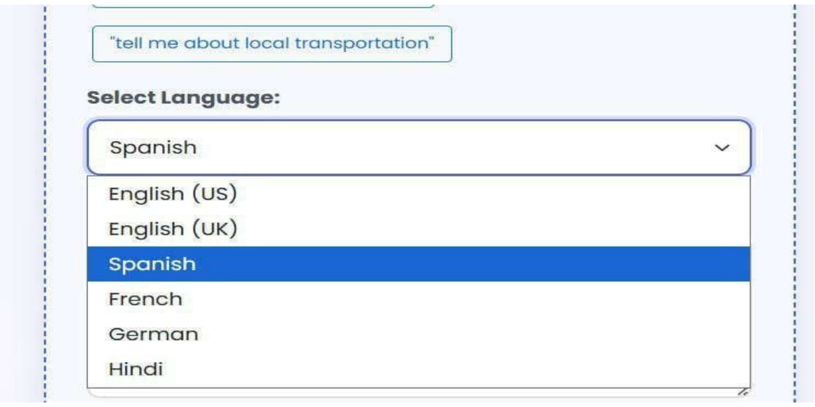
The effectiveness of the translation module is shown by BLEU scores. BLEU evaluates word-level alignment between model predictions and human reference translations. Evaluation of a multilingual speech generation system requires complementary metrics that measure transcription accuracy, translation fidelity, and speech synthesis quality. For ASR, Word Error Rate (WER) is calculated across multiple ac- cents and dialects to ensure real-world robustness. Experiments demonstrate that Whisper-small achieves competitive WER even in noisy conditions, particularly when combined with data augmentation and post-processing filters.



Enter Your Text
Try these examples:
"getmany local bus routs in germany"
"how is the whether in berlin next week"
"best resturants near munich center"
"apply for univercity in germany"
"cheep hotels in frankfurt for turists"
"waht time does the museum open"
"where can i find good food here"
"tell me about local transportation"

Select Language: Spanish

Fig. 3: Example texts to enter the input



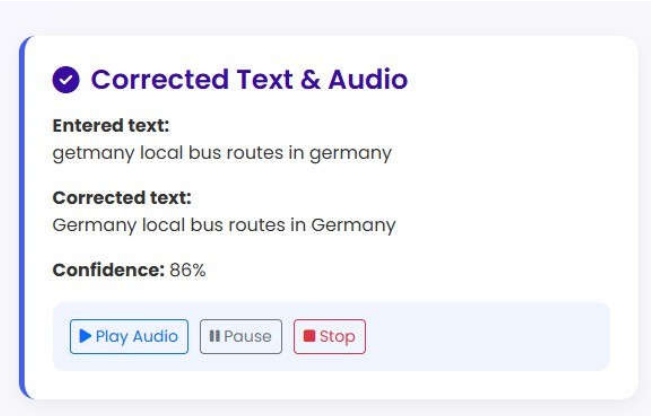
Select Language: Spanish

Fig. 4: Selection of a Languages

Comparative studies show improved accuracy on Indic languages after fine- tuning with region-specific datasets. BLEU score evaluation remains a key metric, but additional human evaluations are performed. Human annotators rate translations based on correctness, fluency, clarity, and cultural appropriateness. Synthesis quality is evaluated using MOS (Mean Opinion Score) ratings. Analysis shows that prosody and emotion modeling significantly enhance perceived naturalness. Entity extraction further reduces mistranslations of technical terms.

7. ADVANCED ARCHITECTURE AND KNOWLEDGE INTEGRATION

The advanced architecture illustrated in integrates knowledge graphs to enhance semantic understanding during translation. Knowledge graphs enable translation engines to map entity relationships, resolve ambiguities, and generate semantically rich translations. The TTS module benefits from these embedding's by adjusting stress and pitch for important terms. This improves user experience, particularly in educational or informational scenarios.



Corrected Text & Audio

Entered text:
getmany local bus routes in germany

Corrected text:
Germany local bus routes in Germany

Confidence: 86%

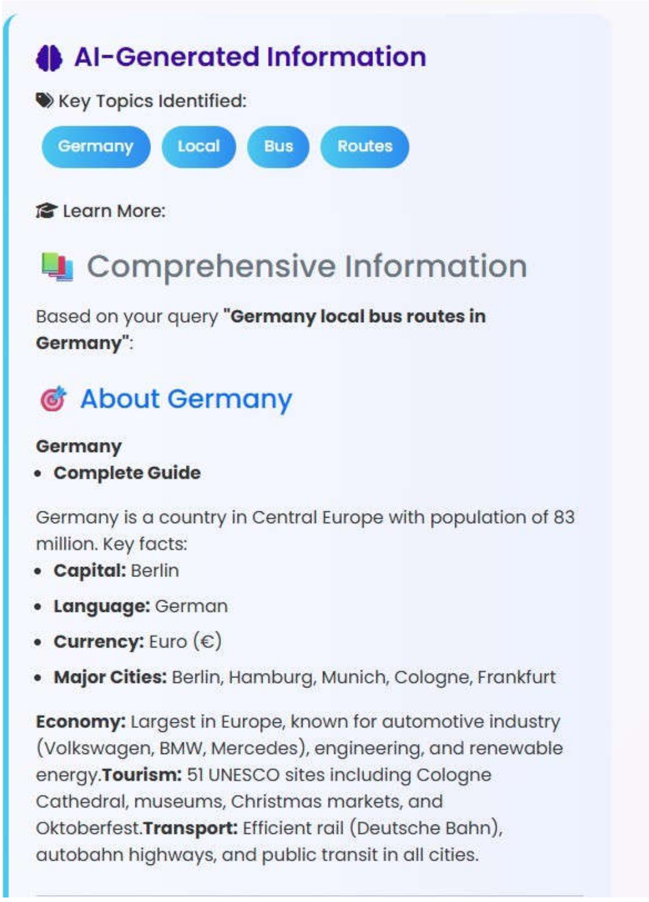
Fig. 5: Corrected text

Advanced multilingual systems often require deeper semantic understanding to translate contextual phrases effectively. Knowledge graphs help disambiguate polysemous words, domain-specific expressions, and culturally sensitive terms. Entity linking improves translation accuracy and ensures the generated speech conveys correct meaning. Within TTS, knowledge integration assists prosody shaping, improving emphasis and clarity for key content.

8. DISCUSSION

The proposed multilingual speech system demonstrates several strengths. Its modular micro service design ensures flexibility, allowing developers to replace or upgrade components independently. The pipeline's strong performance across ASR, NMT, and TTS evaluations indicates that the integration of information extraction and prosody modeling substantially improves translation fidelity and speech quality.

However, challenges remain. Low-resource languages continue to suffer from limited training material, leading to higher WER and lower BLEU scores. Accent variability also presents difficulties. While augmentation techniques improve robustness, they do not fully offset the lack of balanced datasets. High- quality models like VITS require significant GPU resources, making them less suitable for edge devices unless quantized or distilled.



🧠 **AI-Generated Information**

🏷️ Key Topics Identified:

Germany Local Bus Routes

🎓 Learn More:

📚 Comprehensive Information

Based on your query **"Germany local bus routes in Germany"**:

🎯 About Germany

Germany

- **Complete Guide**

Germany is a country in Central Europe with population of 83 million. Key facts:

- **Capital:** Berlin
- **Language:** German
- **Currency:** Euro (€)
- **Major Cities:** Berlin, Hamburg, Munich, Cologne, Frankfurt

Economy: Largest in Europe, known for automotive industry (Volkswagen, BMW, Mercedes), engineering, and renewable energy.**Tourism:** 51 UNESCO sites including Cologne Cathedral, museums, Christmas markets, and Oktoberfest.**Transport:** Efficient rail (Deutsche Bahn), autobahn highways, and public transit in all cities.

Fig. 6: Information Generated

Despite these limitations, the system offers strong potential for practical deployment in educational tools, accessibility platforms, call center automation, and multilingual assistants. Future enhancements may include end-to-end speech-to-speech learning, adaptive prosody based on emotional context, and transformer-based universal speech encoders. Ethical considerations such as deep fake prevention and watermarking must also guide ongoing development.

9. CONCLUSION

This paper presented a comprehensive and modular multilingual Text-to-Speech (TTS) architecture that integrates Automatic Speech Recognition, information extraction, Neural Machine Translation, and expressive speech synthesis into a unified pipeline. Unlike conventional TTS systems that operate in isolation, the proposed architecture demonstrates the benefits of combining multiple deep learning components to achieve context-aware, semantically aligned, and natural speech output across different languages. Through a layered approach, the system demonstrates how linguistic understanding can be enhanced using knowledge integration, entity recognition, and prosody modeling, ultimately generating speech that more closely reflects human-like communication.

The experiments and architectural analysis highlight that multilingual TTS is no longer merely a synthesis challenge—it is a complex pipeline requiring collaboration between perceptual modeling, semantic reasoning, and acoustic generation. By incorporating language identification, advanced translation mechanisms, and prosody conditioning, the system achieves higher robustness and preserves the meaning, emphasis, and sentiment of the source content. The modularity of each subsystem further allows effective scaling, fine-tuning, and replacement, enabling researchers and developers to adapt the framework for new languages, specialized domains, and diverse acoustic conditions.

Overall, the work contributes toward building accessible and intelligent speech interfaces capable of serving global users. The presented architecture offers strong potential for practical deployment in education, healthcare assistance, multi- lingual digital services, and accessibility platforms for visually impaired users. Future research may extend this work by exploring end-to-end speech-to-speech translation systems, emotional awareness models, adaptive prosody learning using reinforcement methods, and advanced knowledge graph integration for deeper semantic fidelity. As speech technology continues to evolve, systems like the one proposed here pave the path toward more inclusive, natural, and contextually adaptive human–machine communication.

REFERENCES

[1] Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[2] Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis." Advances in neural information processing systems 33 (2020): 17022-17033.

[3] Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. "Fastspeech 2: Fast and high-quality end-to-end text to speech." arXiv preprint arXiv:2006.04558 (2020).

[4] Wu, Anne, Changhan Wang, Juan Pino, and Jiatao Gu. "Self-supervised representations improve end-to-end speech translation." arXiv preprint arXiv:2006.12124 (2020).

[5] Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." In International Conference on Machine Learning, pp. 5530-5540. PMLR, 2021.

[6] Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines et al. "Beyond english-centric multilingual machine translation." Journal of Machine Learning Research 22, no. 107 (2021): 1-48.

[7] Babu, Arun, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh et al. "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv preprint arXiv:2111.09296 (2021).

[8] Wang, Changhan, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation." arXiv preprint arXiv:2101.00390 (2021).

[9] Ye, Rong, Mingxuan Wang, and Lei Li. "End-to-end speech translation via cross-modal progressive training." arXiv preprint arXiv:2104.10380 (2021).

[10] Yadav, Hemant, and Sunayana Sitaram. "A survey of multilingual models for automatic speech recognition." arXiv preprint arXiv:2202.12576 (2022).

[11] Zhang, Zi-Qiang, Yan Song, Ming-Hui Wu, Xin Fang, Ian McLoughlin, and Li-Rong Dai. "Cross-lingual self-training to learn multilingual representation for low-resource speech recognition." Circuits, Systems, and Signal Processing 41, no. 12 (2022): 6827-6843.

[12] Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision." In International conference on machine learning, pp. 28492-28518. PMLR, 2023.

[13] Saeki, Takaaki, Heiga Zen, Zhehuai Chen, Nobuyuki Morioka, Gary Wang, Yu Zhang, Ankur Bapna, Andrew Rosenberg, and Bhuvana Ramabhadran. "Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech." In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

[14] Duquenne, Paul-Ambroise, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswami, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. "Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16251-16269. 2023.

[15] Sethiya, Nivedita, and Chandresh Kumar Maurya. "End-to-end speech-to-text translation: A survey." Computer Speech & Language 90 (2025): 101751.

[16] Trenton, Maelle. "A Comprehensive Survey on Multilingual and Multimodal Automatic Speech Recognition Systems." Journal of Computer Technology and Software 4, no. 10 (2025).

[17] Ahlawat, Harsh, Naveen Aggarwal, and Deepti Gupta. "Automatic Speech Recognition: A survey of deep learning techniques and approaches." International Journal of Cognitive Computing in Engineering (2025).