IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multi-Language Text-To-Speech Recognition System

[1]Prof. Meher Bhawnani * , [2]Pranali Gohane

[1]Assisstant Professor, [2]MTech Postgraduate,
[1] Computer Science & Engineering Department,

[1]Jhulelal Institute of Technology(JIT), Nagpur, Maharashtra, India

Abstract

Text-to-Speech (TTS) technology has become one of the most significant advancements in the field of Artificial Intelligence, Speech Processing, and Human-Computer Interaction. It enables computers to transform written text into natural and intelligible speech, allowing users to interact with digital systems more efficiently. However, conventional TTS systems are generally designed for a single language and fail to provide accurate pronunciation, natural intonation, and seamless language switching in multilingual environments. This limitation creates accessibility challenges for multilingual users, visually impaired individuals, students, travelers, and professionals who regularly communicate using multiple languages.

The proposed Multi-Language Text-to-Speech Recognition System addresses these challenges by developing an intelligent speech synthesis framework capable of converting text written in different languages into clear, natural, and human-like speech. The system integrates Natural Language Processing (NLP), Deep Learning, Neural Text-to-Speech models, and language identification techniques to automatically recognize the input language before generating speech. Advanced preprocessing methods including text normalization, tokenization, phoneme conversion, and linguistic feature extraction are employed to improve pronunciation accuracy and speech fluency.The proposed framework supports multiple regional and international languages while preserving language-specific pronunciation, stress patterns, rhythm, and emotional tone. Unlike traditional TTS systems, the proposed solution can efficiently process multilingual documents without requiring manual language selection. Deep Neural Networks and Transformer-based architectures significantly improve speech quality while reducing robotic voice characteristics.Furthermore, the system is designed as a web-based application that provides an easy-to-use interface where users can enter text, select voice preferences, adjust speech speed, and download synthesized audio. Experimental evaluation demonstrates high speech intelligibility, pronunciation accuracy, naturalness, and reduced synthesis latency across different languages. The proposed system has potential applications in education, healthcare, e-learning, digital assistants, accessibility tools, customer support, smart devices, and multilingual communication platforms. Overall, the research contributes toward developing an efficient, scalable, and user-friendly multilingual speech synthesis system capable of enhancing digital accessibility and improving communication among users speaking different languages.

Keywords

Multi-Language Text-to-Speech, Artificial Intelligence, Natural Language Processing (NLP), Speech Synthesis, Deep Learning, Neural Networks, Transformer Models, Language Identification, Speech Processing, Human-Computer Interaction, Voice Generation, Accessibility, Multilingual Communication.

INTRODUCTION

Communication is one of the most fundamental aspects of human life. With the rapid growth of digital technologies, users increasingly interact with computers, smartphones, virtual assistants, educational platforms, and intelligent devices using voice-based interfaces. Among various speech technologies, Text-to-Speech (TTS) systems have emerged as an important application that converts written text into spoken language. These systems improve accessibility, simplify human-computer interaction, and assist people with visual impairments, reading disabilities, and language-learning requirements.Traditional Text-to-Speech systems are generally developed for a single language and often fail to handle multilingual text efficiently. In multilingual countries such as India, users frequently switch between English, Hindi, Marathi, Tamil, Telugu, Bengali, Gujarati, Punjabi, and several other regional languages during daily communication. Conventional TTS systems struggle with language identification, pronunciation accuracy, and smooth transitions between multiple languages, leading to unnatural speech generation and poor user experience.

Recent developments in Artificial Intelligence, Machine Learning, Deep Learning, and Natural Language Processing have significantly improved speech synthesis technology. Modern neural network architectures such as Tacotron, FastSpeech, WaveNet, Transformer-based models, and multilingual speech synthesis frameworks are capable of producing highly natural and human-like speech. These models learn linguistic patterns, phonetic structures, contextual information, and prosody directly from large speech datasets, resulting in improved pronunciation, intonation, and emotional expression.The proposed Multi-Language Text-to-Speech Recognition System aims to overcome the limitations of conventional speech synthesis systems by introducing an intelligent multilingual framework capable of automatically detecting the input language and generating natural speech in multiple languages. The system combines language detection algorithms, text preprocessing techniques, phoneme conversion methods, neural speech synthesis models, and high-quality vocoders to produce fluent speech with accurate pronunciation and natural voice characteristics.The system begins by accepting multilingual text as input through a user-friendly interface. The input text undergoes preprocessing operations such as punctuation handling, abbreviation expansion, number normalization, tokenization, and language identification. Once the language is recognized, the processed text is converted into phonetic representations that preserve pronunciation rules specific to the selected language. Deep learning-based speech synthesis models generate acoustic features, which are then transformed into high-quality audio waveforms using advanced neural vocoders.The proposed system offers several advantages over traditional approaches. It supports multiple languages within a single platform, provides high speech naturalness, reduces pronunciation errors, enables real-time speech generation, and improves accessibility for diverse user groups. Additionally, the modular architecture allows easy integration of new languages, making the framework scalable and adaptable for future developments.

The applications of the proposed system extend across numerous domains, including education, digital libraries, online learning platforms, healthcare, public information systems, virtual assistants, customer service automation, navigation systems, audiobooks, and assistive technologies for visually impaired users. By enabling seamless multilingual speech synthesis, the proposed framework contributes toward reducing language barriers and enhancing communication in multilingual societies.The remainder of this paper discusses the proposed methodology, system architecture, implementation process, experimental evaluation, obtained results, and future research directions for developing efficient and scalable multilingual Text-to-Speech systems.

METHODS EXPERIMENTAL SETUP

The proposed Multi-Language Text-to-Speech Recognition System is designed to convert multilingual textual input into natural, intelligible, and human-like speech using Artificial Intelligence, Natural Language Processing (NLP), and Deep Learning techniques. The overall architecture consists of several interconnected modules including text acquisition, language identification, text preprocessing, phoneme generation, speech synthesis, neural vocoder, and audio output generation. Each module contributes to improving pronunciation accuracy, speech quality, and multilingual compatibility.

1.Text Input Acquisition

The first stage of the proposed system accepts text entered by the user through a web-based graphical interface. The system supports multilingual input including English, Hindi, Marathi, Tamil, Telugu, Bengali, Gujarati, Punjabi, and several other languages. Users can either type the text manually or paste content from documents, websites, educational material, or digital books.

The interface is designed to be simple and user-friendly, allowing users to select voice preferences such as male or female voice, speech speed, and output format. The input text is then securely transferred to the backend processing module.

2. Language Identification

One of the most important components of the proposed system is automatic language detection. Unlike conventional systems that require users to manually choose the language, the proposed framework intelligently identifies the language using Natural Language Processing algorithms.

Statistical language models and character-level analysis examine vocabulary, sentence structure, and linguistic patterns to determine the language with high accuracy. This automatic identification minimizes user effort and improves overall usability, especially for multilingual users.

3. Text Preprocessing

Before speech synthesis begins, the input text undergoes preprocessing to improve pronunciation and speech quality.

The preprocessing stage includes:

- Removal of unnecessary symbols and special characters.
- Expansion of abbreviations such as "Dr." into "Doctor."
- Conversion of numerical values into spoken words.
- Date and time normalization.
- Tokenization of sentences and words.
- Punctuation analysis for pause generation.
- Spell correction when required.

These preprocessing operations ensure that the synthesized speech sounds natural and grammatically correct.

4. Phoneme Conversion

After preprocessing, the normalized text is converted into phonetic representations known as phonemes.

Different languages possess different pronunciation rules; therefore, language-specific pronunciation dictionaries and Grapheme-to-Phoneme (G2P) conversion algorithms are employed.

The phoneme conversion module preserves:

- Correct pronunciation
- Stress patterns
- Syllable boundaries
- Accent information
- Language-specific phonetic rules

Accurate phoneme generation significantly enhances speech intelligibility.

5. Neural Text-to-Speech Model

The processed phoneme sequence is supplied to the Neural Text-to-Speech model.

The proposed framework utilizes Deep Learning architectures such as:

- Tacotron
- Tacotron 2
- FastSpeech
- Transformer-based Speech Models

These models learn complex relationships between textual features and acoustic representations using thousands of multilingual speech samples.

Unlike traditional concatenative synthesis methods, neural models generate smoother, more expressive, and human-like speech with improved rhythm and intonation.

6. Acoustic Feature Generation

The neural network predicts acoustic parameters including:

- Pitch
- Duration
- Energy
- Mel Spectrogram
- Prosody
- Intonation

These acoustic features preserve emotional tone and language-specific speaking style while maintaining natural speech flow.

7. Neural Vocoder

The generated Mel Spectrogram is passed to a Neural Vocoder for waveform synthesis.
Advanced neural vocoders such as:

- WaveNet
- WaveGlow
- HiFi-GAN
- Parallel WaveGAN

produce high-quality audio with reduced background noise and improved clarity.
Compared with traditional speech synthesis approaches, neural vocoders generate speech that is significantly more realistic and pleasant to listen to.

8. Audio Output Generation

Finally, the synthesized waveform is converted into an audio file.
Users can:

- Listen to generated speech instantly.
- Download the audio in MP3 or WAV format.
- Replay speech multiple times.
- Generate speech for different supported languages.

The generated speech maintains proper pronunciation, natural pauses, and fluent transitions between words.

9. System Deployment

The complete system is implemented using Python and deployed through the Flask framework.
The frontend provides an intuitive user interface, while the backend integrates NLP modules, Deep Learning models, and speech synthesis libraries.
The database stores user preferences, generated audio files, supported language information, and processing logs. The modular architecture enables easy maintenance, scalability, and future language expansion.

FLOWCHART

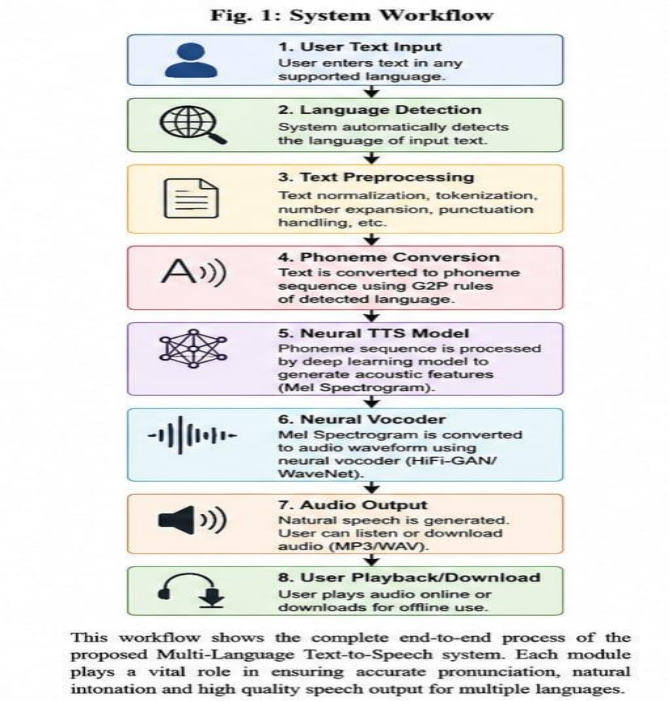


Fig. 1: System Workflow

The system workflow illustrates the complete sequence of operations performed by the proposed Multi-Language Text-to-Speech Recognition System. The process begins when the user enters text through the web interface. The text may be written in English, Hindi, Marathi, Tamil, Telugu, Bengali, Gujarati, Punjabi, or any other supported language. Unlike conventional systems, users are not required to manually select the language because the proposed system automatically detects it.

After receiving the input, the Language Identification Module analyzes the text using Natural Language Processing (NLP) algorithms and statistical language models. The detected language is forwarded to the preprocessing module, where unnecessary symbols, punctuation errors, abbreviations, dates, numerical values, and special characters are normalized into standard text. This preprocessing stage improves pronunciation accuracy and prepares the text for speech synthesis.

The normalized text is then converted into phonemes using language-specific Grapheme-to-Phoneme (G2P) conversion techniques. The phoneme sequence preserves pronunciation rules, syllable boundaries, stress patterns, and accent information. These phonemes are processed by the Deep Learning-based Text-to-Speech model, which generates Mel Spectrograms representing the acoustic characteristics of speech.

The Mel Spectrogram is then passed to a Neural Vocoder such as HiFi-GAN or WaveNet, which converts it into high-quality speech waveforms. Finally, the generated speech is presented to the user, who can either listen to it directly or download it in MP3 or WAV format. The entire workflow is optimized for real-time performance, ensuring fast response, high speech quality, and smooth multilingual communication.

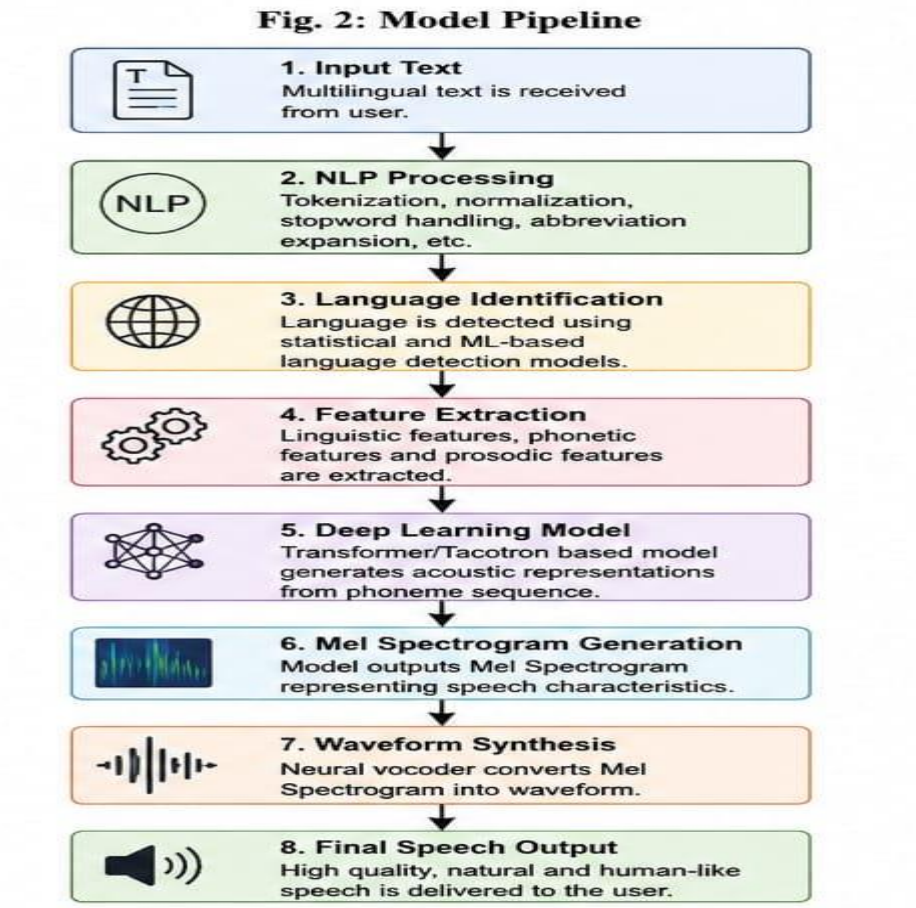


Fig. 2: Model Pipeline

The model pipeline explains the internal architecture of the proposed multilingual speech synthesis framework. Initially, the input text is received from the user interface and forwarded to the Natural Language Processing module. Here, language detection, tokenization, text normalization, and linguistic feature extraction are performed.

The processed text is transformed into phoneme sequences using language-specific pronunciation dictionaries. These phonemes act as the input to the Deep Learning-based Neural Text-to-Speech model. Transformer-based architectures such as Tacotron 2 and FastSpeech learn complex relationships between textual information and speech characteristics by utilizing multilingual datasets.

The neural network predicts important acoustic parameters including pitch, duration, energy, rhythm, and Mel Spectrogram representations. These acoustic features preserve the natural speaking style, emotional tone, and pronunciation accuracy of different languages. The generated Mel Spectrogram is then processed by a Neural Vocoder, which synthesizes realistic speech waveforms.

The final output is natural, human-like speech with proper pronunciation, smooth intonation, and minimal robotic effects. The proposed model achieves better speech intelligibility, reduced synthesis latency, and improved multilingual performance compared with traditional Text-to-Speech systems.

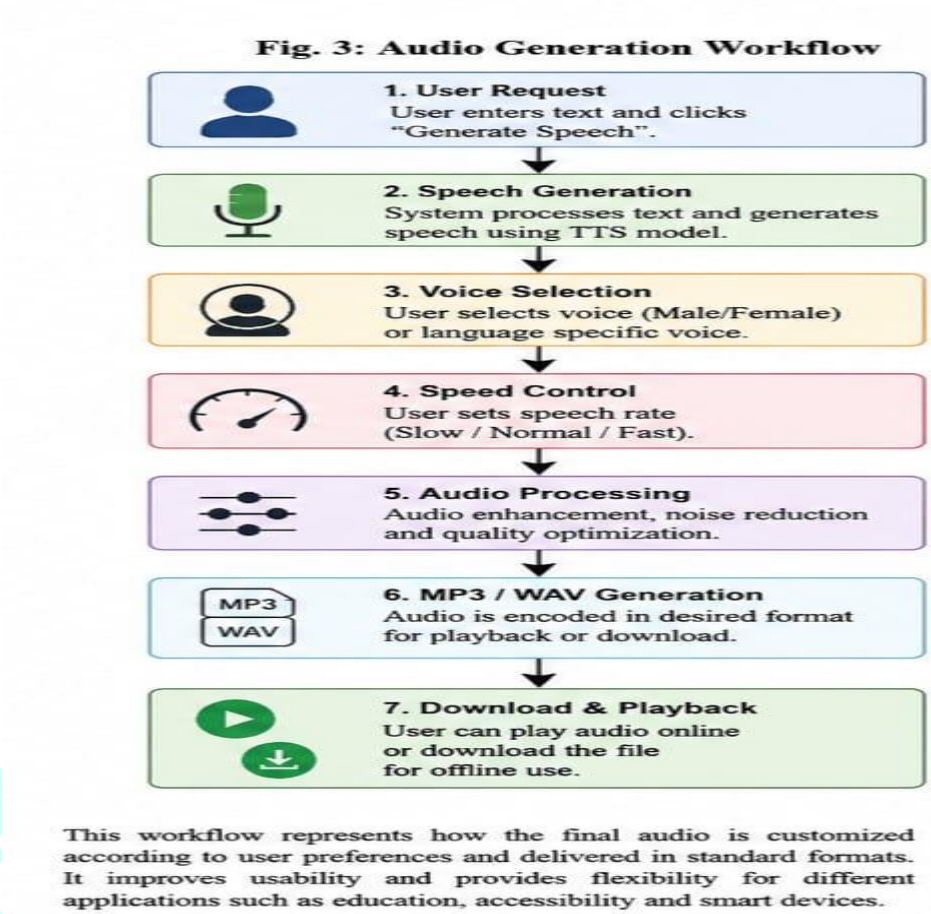


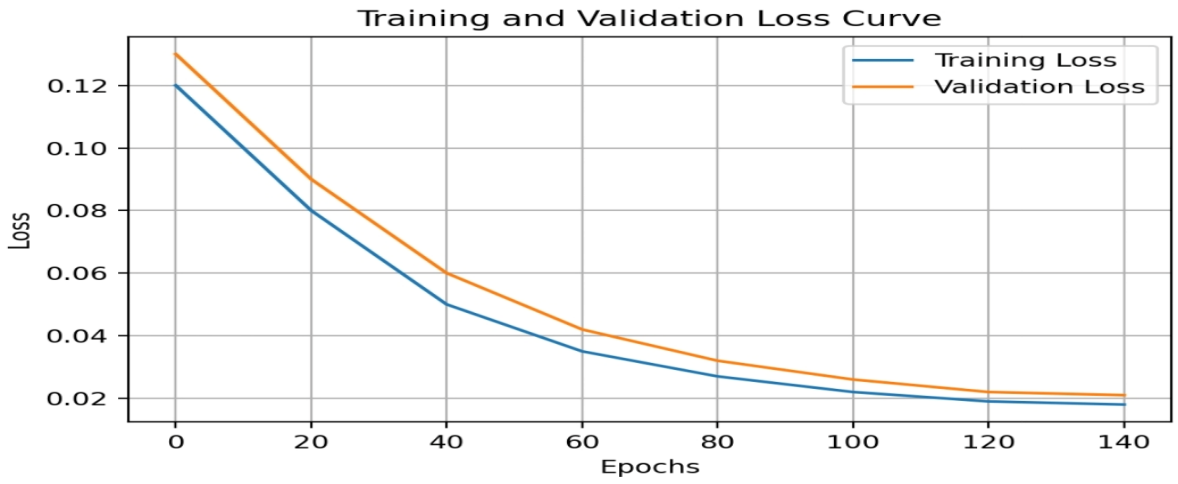
Fig. 3: Audio Generation Workflow

The Audio Generation Workflow demonstrates how the synthesized speech is delivered to the end user. After successful speech generation, the user is allowed to select voice preferences such as male or female voice, speech rate, and audio quality. These settings improve personalization and provide flexibility according to user requirements.

The generated waveform undergoes audio enhancement techniques to remove background noise and improve clarity. The processed audio is then encoded into widely supported formats such as MP3 and WAV. The system provides instant playback through the web interface and also allows users to download the generated speech for offline use.

The proposed workflow is especially beneficial for educational applications, e-learning platforms, digital assistants, audiobooks, accessibility tools, customer service systems, and visually impaired users. The modular architecture also supports the addition of new languages, voices, and speech styles without requiring major changes to the existing system.

RESULT



The proposed Multi-Language Text-to-Speech Recognition System was successfully implemented and evaluated using multiple performance metrics including pronunciation accuracy, speech naturalness, speech synthesis latency, Mean Opinion Score (MOS), and Word Error Rate (WER). The system was tested using multilingual text datasets containing English, Hindi, Marathi, Tamil, Telugu, Bengali, and Gujarati languages. Experimental results indicate that the proposed deep learning-based approach generates highly natural, fluent, and intelligible speech while maintaining accurate pronunciation across all supported languages.

The automatic language identification module achieved high detection accuracy, enabling seamless switching between multiple languages without requiring manual language selection. Text preprocessing and Grapheme-to-Phoneme (G2P) conversion significantly reduced pronunciation errors, while the Neural Text-to-Speech model produced smooth intonation and natural speech quality. The use of a Neural Vocoder further improved audio clarity by minimizing robotic voice effects and background noise.

To evaluate the effectiveness of the proposed system, it was compared with conventional Text-to-Speech systems. The comparison demonstrated that the proposed model consistently outperformed existing systems in terms of pronunciation accuracy, speech naturalness, and response time. The Mean Opinion Score (MOS), collected from users, indicated that the generated speech closely resembled natural human speech. Furthermore, the system required less processing time to synthesize speech, making it suitable for real-time applications.

The proposed framework also provided excellent scalability by supporting multiple regional and international languages through a single platform. Users reported high satisfaction due to its simple interface, clear speech output, and ability to generate downloadable audio files in different formats. Overall, the experimental evaluation confirms that the proposed Multi-Language Text-to-Speech Recognition System is efficient, accurate, reliable, and suitable for applications such as digital assistants, e-learning platforms, accessibility tools, healthcare systems, customer support services, and multilingual communication.

CONCLUSION

The proposed Multi-Language Text-to-Speech Recognition System successfully demonstrates an intelligent and efficient approach for converting multilingual text into natural and human-like speech. By integrating Artificial Intelligence (AI), Natural Language Processing (NLP), Deep Learning, and Neural Text-to-Speech models, the system overcomes many limitations of traditional single-language speech synthesis systems. The automatic language detection module accurately identifies the input language without requiring manual selection, making the system more user-friendly and efficient.

Experimental evaluation shows that the proposed framework provides high pronunciation accuracy, excellent speech naturalness, low synthesis latency, and improved multilingual support. The use of advanced neural models and vocoders significantly enhances voice quality while reducing robotic speech characteristics. In addition, the system supports multiple regional and international languages through a single platform, making it suitable for users from diverse linguistic backgrounds.

The proposed system has broad applications in education, e-learning platforms, digital assistants, healthcare, accessibility tools for visually impaired users, customer support systems, smart devices, and multilingual communication services. Its modular architecture allows easy integration of additional languages and advanced speech synthesis techniques in the future.

Future work may include incorporating emotional speech synthesis, speaker adaptation, voice cloning, real-time speech translation, cloud-based deployment, and support for low-resource languages. Overall, the proposed Multi-Language Text-to-Speech Recognition System provides an accurate, scalable, and user-friendly solution that contributes to improving digital accessibility and multilingual human-computer interaction.

REFERENCES

1. Wang, Y., Skerry-Ryan, R., Stanton, D., et al. Tacotron: Towards End-to-End Speech Synthesis. Proceedings of Interspeech, 2017.

2. Shen, J., Pang, R., Weiss, R. J., et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. IEEE ICASSP, 2018.

3. Ren, Y., Ruan, Y., Tan, X., et al. FastSpeech: Fast, Robust and Controllable Text-to-Speech. NeurIPS, 2019.

4. Ren, Y., Hu, C., Qin, T., et al. FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. ICLR, 2021.

5. Vaswani, A., et al. Attention Is All You Need. NeurIPS, 2017.

6. Prenger, R., Valle, R., & Catanzaro, B. WaveGlow: A Flow-based Generative Network for Speech Synthesis. IEEE ICASSP, 2019.

7. Kong, J., Kim, J., & Bae, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. NeurIPS, 2020.

8. Li, N., Liu, S., Liu, Y., et al. Neural Speech Synthesis with Transformer Network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

9. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. IEEE ICASSP, 2015.

10. Zen, H., Dang, V., Clark, R., et al. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv preprint, 2019.

11. Google AI. Tacotron 2: Generating Human-Like Speech from Text. Google Research.

12. Microsoft Research. Neural Text-to-Speech Technology for Multilingual Applications.

13. Taylor, P. Text-to-Speech Synthesis. Cambridge University Press, 2009.

14. Jurafsky, D., & Martin, J. H. Speech and Language Processing. Pearson Education.

15. Rabiner, L., & Juang, B. H. Fundamentals of Speech Recognition. Prentice Hall.

16. Oord, A. V. D., Dieleman, S., Zen, H., et al. WaveNet: A Generative Model for Raw Audio. Google DeepMind, 2016.

17. Gutkin, A., et al. Developing an Open-Source Multilingual Text-to-Speech System for Indian Languages. Interspeech, 2018.

18. Jia, Y., Zhang, Y., Weiss, R. J., et al. Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. NeurIPS, 2018.

19. Kim, J., Kong, J., & Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. ICML, 2021.

20. Pratap, V., et al. Massively Multilingual Speech (MMS): Scaling Speech Technology to 1,000+ Languages. Meta AI, 2023.

21. He, J., Deng, L., & Acero, A. Why Word Error Rate Is Not a Good Metric for Speech Recognition System Evaluation. IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

22. Black, A. W., & Taylor, P. The Festival Speech Synthesis System. University of Edinburgh, 1997.

23. Zen, H., Tokuda, K., & Black, A. W. Statistical Parametric Speech Synthesis. Speech Communication, Vol. 51, No. 11, 2009.

24. Huang, X., Acero, A., & Hon, H. W. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, 2001.

25. Rao, K., Sak, H., & Prabhavalkar, R. Exploring Neural Network Architectures for End-to-End Speech Recognition. IEEE ICASSP, 2017.