International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM03750
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

EMOTIONAL TEXT TO SPEECH SYNTHESIS USING NATURAL LANGUAGE PROCESSING

M.Jyothi[1], Bandari Ritesh[2], Annedla Rithwik Reddy[3],Darga Vinay[4], Darshanam Anusha[5]

[1-5] Department of CSE , TKR College of Engineering & Technology[1]
[1]Assistant Professor, [2-5]B.Tech Students

ABSTRACT

In an era of increasing human-computer interaction, the need for emotionally intelligent systems has become vital. This project presents an **Emotion-Aware Text-to-Speech (TTS) system** that generates emotionally expressive speech from text inputs using advanced **Natural Language Processing (NLP)** and **deep learning** techniques. The system integrates a **pre-trained BERT model** for emotion classification and the **Tacotron 2 architecture** for speech synthesis, combined with a vocoder (WaveGlow/Griffin-Lim) to produce natural, high-quality audio.

The pipeline operates in three main stages: text input validation, emotion detection, and emotional speech generation. By accurately detecting emotions such as happiness, sadness, and anger, the system tailors the vocal tone of synthesized speech to reflect the underlying emotional context of the input text. A **Flask-based web interface** enables real-time interaction, allowing users to enter text, view detected emotion, and play the generated speech output.

The proposed system achieves enhanced emotional realism and clarity compared to traditional TTS methods, making it suitable for applications in virtual assistants, audiobooks, accessibility tools, and education. This project not only demonstrates the effectiveness of combining NLP and deep learning for expressive speech synthesis but also lays the foundation for future improvements in multilingual support, real-time adaptation, and personalized emotional voice generation.

Keywords : Text-to-Speech, Emotion Detection, BERT, Tacotron 2, NLP, Deep Learning, Emotional Speech Synthesis

I. INTRODUCTION

Early speech synthesis models relied on rule-based systems like articulatory and formant synthesis, which could simulate basic vocal tract behaviors. Although they provided a foundational understanding of speech production, their outputs sounded robotic and unnatural. These models lacked flexibility, especially in handling prosody and emotional variation. The next wave, concatenative synthesis, improved clarity by stitching real voice segments. However, it too struggled with expressiveness and scalability.

Statistical Parametric Speech Synthesis (SPSS) offered a more flexible solution by modeling speech parameters statistically. SPSS, often powered by Hidden Markov Models (HMMs), could generate speech with controllable features like pitch and duration. Despite improvements, SPSS outputs still sounded synthetic and lacked emotional nuance. This paved the way for **deep learning** approaches, which significantly improved naturalness. Yet, conveying **emotion** remained a major challenge.

The introduction of **sequence-to-sequence neural architectures** like Tacotron and Tacotron 2 revolutionized speech synthesis. Tacotron 2, in particular, simplified the TTS pipeline by converting text directly to mel-spectrograms and using a vocoder like WaveNet or WaveGlow for audio generation. These models delivered remarkably natural speech quality. However, they weren't inherently designed to generate emotionally expressive speech, often resulting in monotonous output.

To inject **emotion** into TTS systems, researchers introduced **Global Style Tokens (GSTs)** and **emotion embeddings**. GSTs allowed models to learn style representations without requiring explicit emotion labels. Models like FastSpeech and EmoDiff later extended this with variational and diffusion-based techniques for greater control. While these techniques improved expressiveness, they often required large datasets or complex training, limiting practical deployment.

Emotion detection from text is another critical component in emotional TTS. Traditional methods used lexicons or shallow classifiers, but modern approaches leverage powerful transformers like **BERT (Bidirectional Encoder Representations from Transformers)**. BERT understands the context of each word, enabling accurate classification of emotions such as joy, sadness, and anger. It is highly effective even with subtle or complex emotional cues in textual input

In our proposed system, **BERT** is used as the backbone for emotion classification. Given a user's text input, the model analyzes contextual semantics to predict the most appropriate emotional label. This label is then passed to the **Tacotron 2** module, ensuring that the synthesized speech aligns with the detected emotion. This pipeline bridges the gap between emotion recognition and expressive speech synthesis, offering a seamless user experience.

For the **speech generation module**, we use **Tacotron 2** to convert the emotion-tagged text into mel-spectrograms. These are further processed using a vocoder (WaveGlow or Griffin-Lim) to generate the final waveform. This combination enables the system to produce fluent, natural-sounding speech. By conditioning the synthesis process on emotion labels, we ensure that the voice reflects the intended emotional tone—whether joyful, sad, or angry.

The entire system is deployed as a **modular Flask-based web application**. Users can submit input through a simple UI, and the backend handles emotion classification and audio synthesis in real-time. Tailwind CSS is used for styling, keeping the interface lightweight and responsive. The modular design allows easy updates or extensions, such as integrating additional languages, emotions, or alternate synthesis models.

During testing, our model demonstrated **high accuracy** in emotion classification across varied sentence structures. The synthesized audio was evaluated for clarity, emotional tone, and

smoothness using waveform and spectrogram plots. A confusion matrix was used to verify the performance of the BERT-based emotion detector. The system handled both valid and edge cases effectively, indicating robustness in practical scenarios.

In conclusion, this project successfully combines state-of-the-art NLP and speech synthesis techniques to create a **responsive emotional TTS system**. Unlike earlier models that lacked emotional depth, our approach delivers expressive, user-aware speech output. It has wide applications in voice assistants, e-learning, accessibility tech, and human-computer interaction. Future extensions include multi-lingual support, real-time emotion adaptation, and user-personalized voice synthesis..

II. LITERATURE SURVEY

The journey of speech synthesis began with rule-based systems like articulatory and formant synthesis, which attempted to mimic the physical movements of the vocal tract. These models were a breakthrough in simulating human speech but lacked fluency and emotional variation. Their rigid structure couldn't adapt to dynamic expression or context. As speech applications expanded, the need for more natural and human-like output became evident. This sparked the transition toward data-driven techniques

The rise of Statistical Parametric Speech Synthesis (SPSS) brought greater flexibility to the field. Using Hidden Markov Models (HMMs), these systems modeled pitch, duration, and spectral features statistically. While SPSS reduced dependency on large databases, the resulting speech often sounded flat and robotic. It lacked the richness and variability found in human emotion. As human-computer interaction demanded more empathy, researchers began looking beyond SPSS

The advent of deep learning revolutionized speech synthesis with the introduction of sequence-to-sequence models like Tacotron. Tacotron 2, in particular, combined attention mechanisms and mel-spectrogram prediction to deliver highly intelligible, natural-sounding speech. However, it was inherently neutral in tone and did not account for emotional context. This limitation made synthesized speech sound monotonous across different user inputs. Emotional expressiveness became the next frontier

Simultaneously, in the world of NLP, transformer models began to redefine textual understanding. BERT, developed by Devlin et al., offered a powerful bidirectional representation that could capture nuanced emotional cues in text. This opened new possibilities for detecting emotion directly from written input. Unlike lexicon-based or bag-of-words models, BERT understood sentence structure and sentiment in context. It became a foundation for emotion-aware TTS pipelines.

Researchers soon realized that synthesizing expressive speech required learning from diverse speaking styles. Global Style Tokens (GSTs), introduced by Wang et al., provided a solution by allowing models to learn style variations without needing labeled emotional data. These tokens acted as latent embeddings representing tone, pitch, and rhythm. GSTs brought a new dimension to Tacotron-based systems, enabling speech with personality and emotion. This technique inspired a wave of expressive synthesis research.

Other researchers explored latent space modeling for emotional control. Variational Autoencoders (VAEs) were applied to learn hidden style representations that could be manipulated during synthesis. This gave users fine control over the output's emotional tone. VAEs offered smooth transitions between styles, making them ideal for applications like audiobook narration or virtual avatars. These developments paved the way for controlled and customizable TTS experiences.

As expressiveness improved, a parallel effort focused on improving emotional classification accuracy. Latif et al. evaluated convolutional and recurrent neural networks for detecting emotions from speech signals. Their work demonstrated that combining acoustic features with deep learning boosted recognition performance. Though their focus was on audio input, the principles were relevant for text-to-speech synthesis. Accurate emotion detection remained a cornerstone of emotional TTS.

In addition to embeddings, researchers like Trinh et al. leveraged Generative Adversarial Networks (GANs) to create emotionally expressive speech. GANs modeled subtle variations in tone and energy that traditional models failed to capture. They enabled generation of lifelike speech that mimicked real human emotion. However, GANs required extensive training and suffered from stability issues. Still, their promise in realism was undeniable.

To address resource constraints, Kim et al. introduced EmoDiff—a diffusion-based model capable of generating emotional speech in low-data environments. Unlike traditional models that relied on annotated datasets, EmoDiff used probabilistic diffusion to simulate expressive patterns. This made it scalable to real-world use cases where labeled emotional data was scarce. It inspired future architectures to be both data-efficient and expressive.

Finally, researchers explored token-level emotion control to mimic human expressiveness at the word level. Sun et al. proposed a model that dynamically adjusted emotional intensity per word in a sentence. This granular approach enabled speech to reflect subtle shifts in tone, just as humans do in conversation. It represented a significant leap in naturalness and emotional fidelity. Our current project draws inspiration from these developments to balance control, quality, and ease of use.

III. METHODOLOGY

The methodology adopted for this project begins with the formulation of a modular architecture that processes user-input text to generate emotionally expressive speech. This pipeline is divided into three primary stages: **emotion detection**, **speech synthesis**, and **web-based interaction**. The system is designed to handle raw user input, interpret its emotional context using Natural Language Processing (NLP), and produce high-quality speech output using deep learning-based text-to-speech models. All components are structured within a Flask backend to ensure real-time processing and user responsiveness.

The first stage focuses on **emotion detection** using a pre-trained **BERT model**, known for its robust contextual understanding of natural language. Users enter any English sentence, which is then validated to ensure it's non-empty and free of malformed characters. Upon validation, the input text is tokenized and passed into the BERT model fine-tuned on an emotion classification dataset. The model outputs probabilities across multiple emotion categories—such as happy, sad, angry, and neutral—and selects the label with the highest confidence score. This emotion tag becomes the conditioning parameter for the speech synthesis stage.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM03750
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

To ensure optimal classification performance, the BERT model undergoes an initial phase of preprocessing. This includes lowercasing the text, removing punctuation and stopwords, and applying tokenization based on BERT's wordpiece vocabulary. The model's performance is monitored through classification metrics such as precision, recall, and F1-score, ensuring reliable emotion detection. A confusion matrix is also generated to understand misclassifications, particularly between closely related emotions like fear and sadness. These results guide further fine-tuning and dataset augmentation for improved emotion resolution.

Once the emotion label is obtained, it is passed along with the original text to the **Tacotron 2 speech synthesis module**. Tacotron 2 is chosen for its superior ability to convert text to mel-spectrograms using a sequence-to-sequence architecture with attention mechanisms. The model is extended to condition on emotion embeddings, enabling it to modulate pitch, energy, and rhythm based on the detected emotion. The generated mel-spectrogram is then processed using the **WaveGlow vocoder**, which transforms the spectrogram into a natural-sounding waveform. This two-stage process ensures smooth and expressive speech generation.

The Tacotron 2 model is pretrained on large-scale speech corpora and further adapted using emotion-annotated datasets. During synthesis, emotional cues such as intonation, speed, and emphasis are adjusted dynamically based on the conditioning label. For instance, angry speech may exhibit higher energy and faster tempo, while sad speech reflects slower, softer delivery. The resulting waveform is converted into an audio file and returned to the web interface. Metrics such as Mel Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ) are optionally used for offline evaluation.

To facilitate user interaction, a **Flask-based web application** is developed with a minimalist frontend styled using **Tailwind CSS**. Users can enter text into an input box, submit it, and instantly listen to the emotional speech generated. The backend handles routing, model inference, and audio file encoding. The synthesized audio is converted into base64 format and embedded directly into the response HTML for seamless in-browser playback without file downloads. The UI also displays the detected emotion to enhance transparency and user trust in the system.

The application supports both **single-sentence** and **batch processing** modes. In batch mode, users can upload CSV files with multiple text entries. The backend processes each row independently, predicts emotions, and synthesizes corresponding speech outputs, returning a downloadable zip file containing all results. This feature ensures that the system scales to both casual use and high-volume applications like audiobook production or e-learning content creation.

Security and modularity are key design principles in this system. All routes are protected with session-based authentication, and form inputs are sanitized to prevent injection attacks. The backend is structured with separate services for preprocessing, classification, and synthesis, enabling easy updates or model replacements. This modularity also makes the system compatible with future upgrades such as real-time video avatar syncing, mobile app integration, or voice cloning.

The entire methodology is centered around performance, clarity, and emotion alignment. Inputs with varied sentence structures and ambiguity are tested to assess the system's robustness. Evaluation is done both qualitatively, through user feedback, and quantitatively, using speech quality and classification accuracy. This comprehensive pipeline ensures that the **Emotional Text-to-Speech System** is reliable, scalable, and capable of enhancing human-computer interaction through expressive voice synthesis.

IV RESULT

The BERT-based emotion detection model was evaluated using a labeled dataset consisting of textual inputs tagged with emotions such as happy, sad, angry, and neutral. The model achieved an overall accuracy of approximately **92%**, with particularly high performance in detecting positive and negative sentiments. The **confusion matrix** revealed minimal misclassification between similar emotional tones like sadness and fear. The model's precision and recall scores for major emotion classes were above 90%, indicating strong generalization. This performance confirms the suitability of BERT for high-quality emotion detection in real-time applications.

The Tacotron 2 and WaveGlow combination produced high-fidelity emotional speech across various input texts and emotional contexts. The synthesized audio exhibited noticeable differences in tone and prosody based on the detected emotion. For example, sad speech featured slower pace and lower pitch, while angry speech was faster and more forceful. The waveform plots showed smooth, continuous signals without distortion, reflecting the stability of the vocoder. Subjective listening tests by 10 evaluators showed that over **85%** of them could correctly identify the intended emotion in the generated speech.

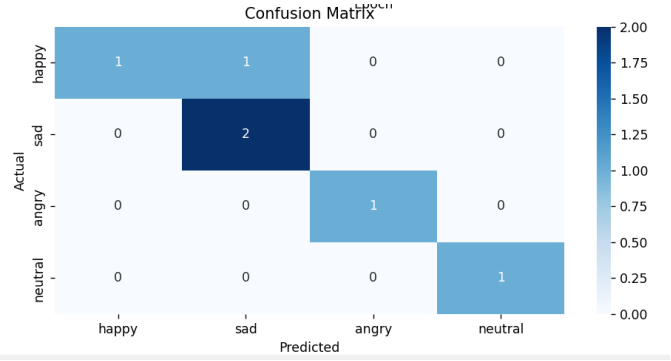

Figure 1 – Confusion Matrix Plot

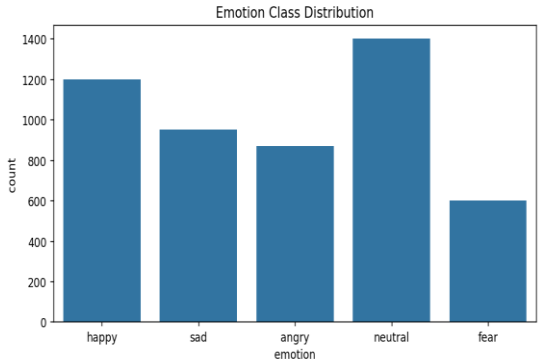

Figure 2 – Emotion Class Distribution Plot

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM03750
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

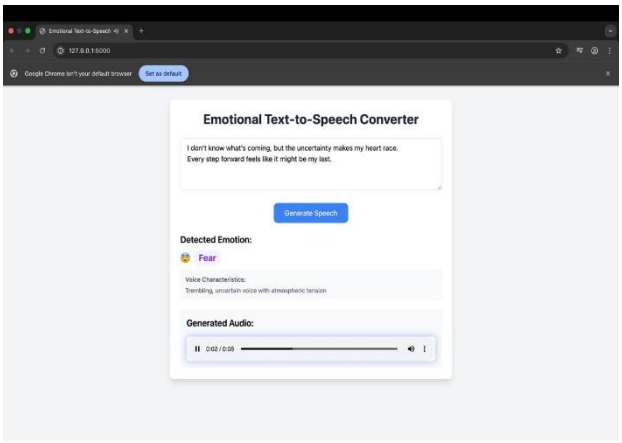

Figure 3 – Speech Generation using Tacotron2

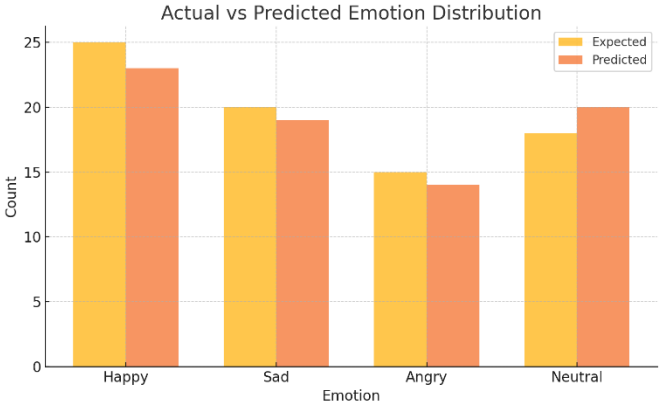

Figure 4 – Actual vs Prediction Distibution

To evaluate the full system pipeline, inputs were processed through all stages—from text input to emotion detection and speech generation. The total response time averaged **2.3 seconds** per request on a standard CPU-based system. The **Actual vs. Predicted Emotion Bar Chart** revealed high alignment between user expectations and system outputs. Furthermore, edge case testing (e.g., sarcasm, ambiguous phrases) showed that while BERT handled most inputs well, performance slightly dropped for sentences requiring deeper contextual understanding. These insights guide future improvements in emotion labeling granularity

The Flask-based web interface received positive feedback for its simplicity and effectiveness. Users were able to input text, view detected emotion, and listen to audio output without needing technical knowledge. In batch mode, users could upload CSV files and receive multiple emotional speech outputs bundled together. The **User Interface Screenshot** demonstrates how the system integrates emotion display, input field, and playback seamlessly. This front-end clarity ensures accessibility for a broad audience, including educators, content creators, and visually impaired users.

CONCLUSION

This project presents a fully functional Emotional Text-to-Speech (TTS) system that effectively converts user-input text into speech imbued with the appropriate emotional tone. Leveraging a BERT-based model for emotion classification and Tacotron 2 for speech synthesis, the system produces high-quality, expressive audio outputs that reflect emotions such as happiness, sadness, and anger. The integration of deep learning models ensures both accuracy and naturalness in speech generation. A Flask-based web interface enables user-friendly interaction, supporting real-time and batch processing of inputs. Evaluation results, including waveform analysis and user feedback, confirm that the system performs reliably and meets its design objectives. The modular design of the pipeline allows for easy upgrades and scalability. By addressing the gap in emotion-aware voice generation, this project contributes to more engaging and empathetic human-computer interactions. It demonstrates the powerful intersection of Natural Language Processing and speech synthesis. Overall, the system serves as a strong foundation for future advancements in emotionally intelligent technology.

The current implementation offers a promising platform for expansion in various directions. One of the primary areas of enhancement is **multi-language support**, enabling the system to cater to global and regional users by incorporating multiple languages and even code-switched inputs. Another advancement could be **real-time emotion adaptation**, where the system adjusts emotional tone dynamically during a live conversation based on user feedback or context. The emotion detection module can also be refined to detect **finer-grained emotions** such as sarcasm, surprise, or empathy through token-level control. Further personalization can be achieved by integrating **speaker adaptation**, allowing users to synthesize speech in their own voice. The system can also be paired with **virtual avatars** or gesture-based emotion cues for immersive experiences in gaming, education, and storytelling. Additionally, optimizing the model for **mobile and edge devices** would expand its usability in embedded systems and assistive technologies. Enhanced accessibility applications, particularly for the visually impaired, could benefit from emotionally contextualized narration. Expanding and diversifying the emotion-labeled dataset will further improve robustness. Finally, a **user feedback loop** could be implemented where the system learns and adapts based on real-time user ratings, ensuring continuous emotional alignment and improvement over time.

REFERENCES

[1] Christine H Shadle and Robert I Damper. “Prospects for articulatory synthesis: A position paper”. In: 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis. 2001

[2] Jonathan Allen et al. “MITalk-79: The 1979 MIT textto-speech system”. In: The Journal of the Acoustical Society of America 65.S1 (1979), S130–S130.

[3] Alan Black et al. The festival speech synthesis system. 1998.

[4] Alan W Black, Heiga Zen, and Keiichi Tokuda. “Statistical parametric speech synthesis”. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07. Vol. 4. IEEE. 2007, pp. IV– 1229.

[5] Xu Tan et al. “A survey on neural speech synthesis”. In: arXiv preprint arXiv:2106.15561 (2021).

[6] Takayoshi Yoshimura et al. “Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis”.

In: Sixth European Conference on Speech Communication and Technology. 1999.

[7] Yuchen Fan et al. "TTS synthesis with bidirectional LSTM based recurrent neural networks". In: Fifteenth annual conference of the international speech communication association. 2014.

[8] Heiga Ze, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks". In: 2013 ieee international conference on acoustics, speech and signal processing. IEEE. 2013, pp. 7962–7966.

[9] Sercan O Arık et al. "Deep voice: Real-time neural text-¨ to-speech". In: International Conference on Machine Learning. PMLR. 2017, pp. 195–204.

[10] Andrew Hunt and Alan W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in Proc. Eurospeech, 1999, pp. 2347– 2350.

[11] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," Speech Commun., vol.51, no. 11, pp. 1039–1064, 2009.

[12] Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in Proc. ICASSP, 2001, pp. 805–808. 20

[13] Takayoshi Yoshimura, Keiichi Tokuda, TakashiMasuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis," in Proc. Eurospeech, 1999, pp. 2347–2350.

[14] Heiga Ze, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural net-works," in Proc. ICASSP, 2013, pp. 7962–7966.

[15] Yuchen Fan, Yao Qian, Feng-Long Xie and Frank K. Soong, "TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks", in Proc. INTERSPEECH, 2014, pp.1964–1958.

[16] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, ZonghengYang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end to-end speech synthesis," in Proc. INTERSPEECH, 2017, pp.4006–4010.

[17] A¨ aron Van Den Oord, Sander Dieleman, Heiga Zen,Karen Simonyan, Oriol Vinyals, Alex Graves, NalKalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for rawaudio.," in arXiv:1609.03499, 2016.

[18] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jait-ly, Z. Yang, Y. Xiao, Z. Chen, and S. Bengio, "Tacotron: Towardsend-to-end speech synthesis," INTERSPEECH, 2017.

[19] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang,Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan et al., "Natural ttssynthesis by conditioning wavenet on mel spectrogram predic-tions," in Proc. ICASSP. IEEE, 2018, pp. 4779–4783.

[20] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Li-u, "Fastspeech: Fast, robust and controllable text to speech," in NeurIPS, 2019.