International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 12, pp 192-198, December-2025

https://doi.org/10.47001/IRJIET/2025.912029

Text To Speech with Emotion Control Using Deep Learning

[1]Manasa S M, [2]Subramanya S Gujjar, [3]Shashank H S, [4]Subhash G K, [5]Vikyath M A

[1]Assistant Professor, Department of IS&E, JNNCE, Shivamogga, Karnataka, India

[2,3,4,5]UG Student, Department of IS&E, JNNCE, Shivamogga, Karnataka, India

***Abstract* - Recent advances in neural Text-to-Speech (TTS) systems have produced speech with high intelligibility and naturalness, yet most deployed systems still sound emotionally neutral. This lack of affective expressiveness limits user engagement and degrades the quality of human– computer interaction, especially in applications such as virtual assistants, audiobooks, education, and accessibility technologies. This work proposes an emotion-infused TTS framework that extends a Tacotron-based sequence-to-sequence architecture with explicit emotion conditioning. The system leverages the Emotional Speech Database (ESD) to model five emotional categories—neutral, happy, angry, sad, and surprise—and incorporates emotion vectors alongside text embeddings in the encoder–decoder pipeline. Mel-spectrograms predicted by the model are converted to waveforms using the Griffin–Lim algorithm. Experimental training on English-emotion subsets of ESD demonstrates stable convergence of mel-spectrogram reconstruction loss and the capability to synthesize perceptually distinct emotional speech, as observed through qualitative waveform and spectrogram analysis. A web-based interface is further developed to enable end-user interaction, allowing text input or file upload with selectable emotional style. The proposed system shows that explicit emotion conditioning can significantly enhance expressiveness of neural TTS without sacrificing intelligibility, and it provides a practical foundation for emotionally aware human–machine communication.**

Keywords: Text-to-Speech, Emotion Control, Deep Learning, Tacotron, Mel-Spectrogram, Speech Synthesis.

I. INTRODUCTION

Neural Text-to-Speech (TTS) has progressed from concatenative and statistical parametric methods to powerful end-to-end architectures capable of producing highly natural speech. Models such as Tacotron, Tacotron 2, Fast Speech, and other sequence-to-sequence frameworks with attention have demonstrated that deep neural networks can learn direct mappings from text to acoustic representations, eliminating the need for complex hand-crafted linguistic front-ends. However, despite these advances, mainstream TTS systems remain predominantly neutral in prosody and affect, which makes synthetic speech sound flat and emotionally unengaging.

In natural human communication, emotion is expressed through systematic variations in pitch, energy, timing, and spectral characteristics. These affective cues are crucial not only for perceived naturalness but also for pragmatic understanding—distinguishing, for example, between sarcasm and sincerity, or urgency and calm. A TTS system that ignores emotional expressiveness can therefore deliver correct words while still failing to convey intended communicative intent.

Embedding emotion into TTS is challenging for several reasons. First, emotion is multidimensional and subjective; categorical labels such as “happy” or “sad” are simplifications of complex affective states. Second, emotional prosody is entangled with speaker identity, linguistic content, and contextual factors, making it difficult to disentangle and control. Third, large, high-quality emotional speech corpora are less common than neutral datasets, limiting data-driven approaches.

This paper addresses these challenges by extending a Tacotron-based architecture with an explicit emotion encoder and conditioning mechanism. Instead of relying on implicit prosodic variation learned from mixed-emotion corpora, the model takes both text and emotion as inputs and learns to generate mel-spectrograms conditioned on a specific emotional target. The Emotional Speech Database (ESD) is used as the primary corpus, providing controlled recordings across five emotion classes. Mel-spectrograms are reconstructed to waveform using the Griffin–Lim algorithm, making the system fully neural in the text-to-spectrogram stage while keeping the vocoder classical and interpretable. The main contributions of this work are:

1. Design of a Tacotron-based TTS architecture with explicit emotion conditioning via a dedicated emotion encoder.
2. A complete preprocessing and training pipeline leveraging ESD for multi-emotion speech synthesis.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 12, pp 192-198, December-2025

https://doi.org/10.47001/IRJIET/2025.912029

3. A qualitative evaluation of multi-emotion synthesis through spectrograms and waveform analysis.
4. Development of a web-based interface enabling users to generate emotionally controlled speech on demand.

II. RELATED WORK

Early TTS systems were based on concatenative synthesis, concatenating prerecorded units such as diphones or syllables. While capable of high-quality output in constrained conditions, they suffered from limited prosodic flexibility and required large, carefully curated databases. Statistical parametric speech synthesis (SPSS) using Hidden Markov Models (HMMs) partially addressed these issues by modeling acoustic parameters statistically, enabling smoother prosodic modifications at the expense of naturalness.

The advent of deep learning led to neural acoustic models, initially using feed forward networks and recurrent neural networks (RNNs) to map linguistic features to acoustic parameters. The breakthrough came with sequence-to-sequence models with attention, notably Tacotron, which demonstrated that a single neural network can learn text-to-spectrogram mappings without explicit alignment. Subsequent refinements in Tacotron 2 and non-autoregressive models such as FastSpeech improved training stability, synthesis speed, and overall naturalness.

Parallel to TTS, Speech Emotion Recognition (SER) research has characterized emotional cues in speech using prosodic, spectral, and linguistic features. SER studies show that emotions such as anger and happiness strongly correlate with changes in fundamental frequency contours, intensity, and speaking rate, while sadness is often associated with lower energy and slower speech. These insights are crucial for designing emotion synthesis models that mimic human-like expressive patterns.

Emotion-aware TTS research has followed several directions:

- Label-based conditioning, where discrete emotion labels are embedded and concatenated with text or acoustic features.
- Style tokens and global style embeddings, where latent vectors capture different prosodic styles, including emotional variations, without explicit labels.
- Variational and generative models, using VAEs or GANs to disentangle style (emotion, speaking style) from content.

While these approaches show promise, many are complex to train or less accessible for direct deployment. The approach in this work deliberately uses a more transparent architecture: a Tacotron-like model with a dedicated emotion encoder, focusing on reproducibility and clarity rather than cutting-edge generative complexity.

III. SYSTEM ARCHITECTURE

3.1 Overall Framework

The proposed system follows a modular pipeline:

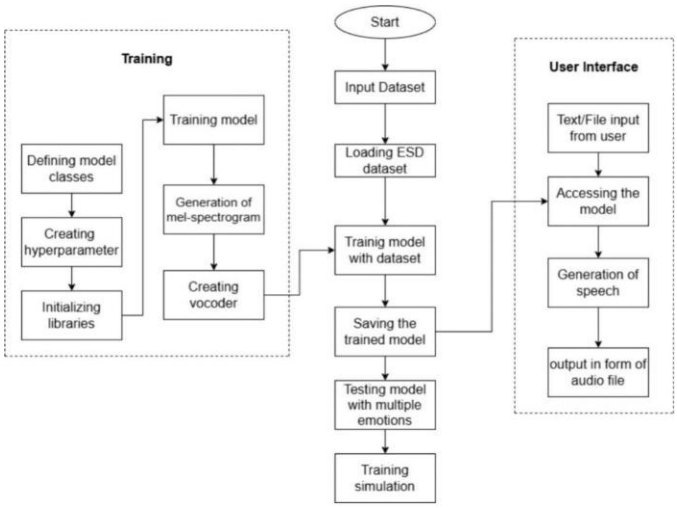

Figure 1: Flow Chart

1. Text Preprocessing and Encoding: As shown in the above figure 1 raw input text is normalized, tokenized, and converted to a sequence of integer token IDs. These tokens are embedded and passed through a bidirectional encoder to capture linguistic context.
2. Emotion Encoding: Emotion labels (neutral, happy, angry, sad, surprise) are mapped to learned embedding vectors representing emotion-specific characteristics. An emotion encoder refines these embeddings to generate compact emotion vectors.
3. Sequence-to-Sequence Decoder with Attention: A Tacotron-style decoder uses attention to align encoded text–emotion representations with output time steps. At each decoding step, the model predicts a frame of the mel spectrogram conditioned on the previous acoustic frame, the current decoder state, and the attention-weighted encoder context.
4. Post-Net and Mel-Spectrogram Refinement: A convolutional post-net refines initial mel-spectrogram predictions, improving spectral detail.
5. Waveform Reconstruction: The Griffin–Lim algorithm is applied to mel-spectrograms (mapped to linear magnitude spectra as needed) to reconstruct time-domain waveforms.
6. Web-Based Inference Interface: A Flask-based web application provides a user-facing interface where text can be entered or files uploaded, and an emotion can be selected to generate downloadable speech audio.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 12, pp 192-198, December-2025

https://doi.org/10.47001/IRJIET/2025.912029

3.2 Dataset and Preprocessing

This work uses the English subset of the Emotional Speech Database ESD, comprising multiple speakers recording phonetically balanced sentences under five emotional conditions: neutral, happy, angry, sad, and surprise. Each recording typically spans 1-3 seconds, producing sufficient coverage of phonetic and prosodic patterns.

Preprocessing steps include:

1. Audio Normalization: Resampling to a fixed sampling rate (e.g., 22.05 kHz). Peak normalization to remove recording-level variations.
2. Text Cleaning and Normalization: Removal of non-linguistic symbols and special characters. Conversion to lowercase and standard punctuation normalization. Mapping to character or phoneme-level tokens, depending on implementation.
3. Feature Extraction: Short-Time Fourier Transform STFT) with fixed window and hop sizes. Mapping magnitude spectra to mel-spectrograms (e.g., 80 mel bins). Log scaling and mean–variance normalization across the training set.
4. Emotion Label Handling: Assigning each utterance a categorical emotion label. Converting labels to one-hot vectors and then to continuous embeddings through a trainable lookup layer.
5. Sequence Alignment and Padding: Padding text and mel-spectrogram sequences to mini-batch compatible lengths. Masking padded regions during loss computation to avoid bias.

3.3 Tacotron-Based Emotion-Conditioned Model

The model maintains the core structure of Tacotron but is extended to incorporate emotion conditioning:

- Text Encoder: An embedding layer maps token IDs to dense vectors, which are passed through convolutional layers and a bidirectional recurrent layer (e.g., BiLSTM or BiGRU. The resulting hidden states capture local and long-range textual dependencies.
- Emotion Encoder: A small feedforward network maps discrete emotion embeddings to a fixed-length emotion vector. This vector is broadcast or concatenated to each encoder time step, enabling the model to condition the entire sequence on the target emotion.
- Attention Mechanism: A location-sensitive attention computes alignment between decoder states and encoder outputs. Because emotion vectors are integrated into encoder outputs, attention implicitly learns emotion-conditioned alignment patterns where necessary.
- Decoder: An autoregressive RNN-based decoder predicts mel-spectrogram frames. During training, teacher forcing is used—ground truth mel frames are fed as inputs. During inference, the decoder uses its own previous predictions.

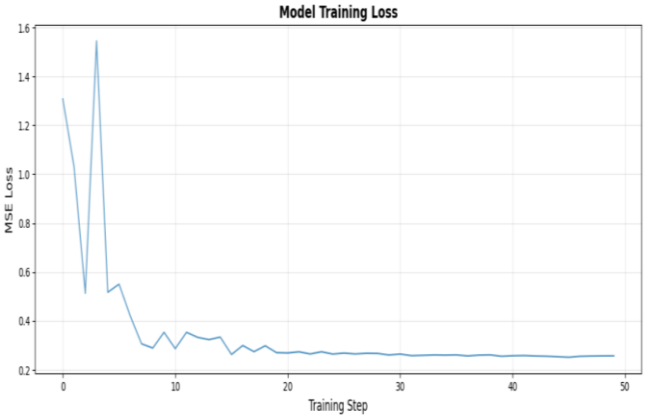

Figure 2: Model training loss

The figure 2 is generated based on the model training, if t denotes the token sequence,$e$the emotion label, and m the mel-spectrogram, the model approximates:

$$P(m \mid t, e) = \prod_{i=1}^{T} P(m_i \mid m_{<i}, t, e)$$

Where m_i is the i-th mel frame and T is the length of the mel sequence.

IV. TRAINING PROCEDURE

4.1 Optimization setup

The model is trained using the Adam optimizer with a typical learning rate on the order of , decayed according to validation performance. Batch sizes are selected based on GPU memory constraints (e.g., 1632 utterances per batch). Gradients are clipped to maintain training stability.

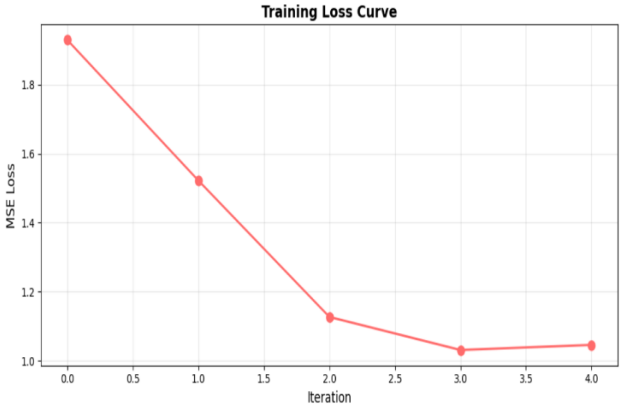

Figure 3: Training loss curve

ISSN (online): 2581-3048
Volume 9, Issue 12, pp 192-198, December-2025
https://doi.org/10.47001/IRJIET/2025.912029

The primary loss function is a combination of:

- Mean Absolute Error MAE) or Mean Squared Error MSE) between predicted and target mel-spectrograms (before and after the post-net).
- Optional stop-token loss predicting the end of the sequence.

The total loss as shown in the figure 3 can be summarized as:

$$\mathcal{L} = \lambda_1 \|\hat{M} - M\|_1 + \lambda_2 \|\hat{M}^{post} - M\|_1 + \lambda_3 \mathcal{L}_{stop}$$

Where $\hat{M}$ is the initial mel prediction, $\hat{M}^{post}$ is post-net output, and M is the ground truth.

4.2 Training Dynamics

Across tens of thousands of iterations, both linear and mel-spectrogram losses decrease monotonically and eventually plateau, indicating stable convergence. The lack of large oscillations or divergence in loss curves suggests:

- Appropriate model capacity relative to dataset size
- Reasonable choice of learning rate and optimizer parameters
- Effective handling of sequence padding and masking

Validation loss is monitored to detect overfitting; early stopping or learning rate adjustments are applied if validation loss stagnates or worsens.

V. WAVEFORM RECONSTRUCTION WITH GRIFFIN–LIM

Since Tacotron predicts mel-spectrogram magnitude values rather than full complex STFTs, a phase recovery method is required to synthesize waveforms. The Griffin–Lim algorithm is a widely used iterative technique for reconstructing phase from magnitude spectra:

1. Initialize complex spectrogram with the predicted magnitude and random phase.
2. Apply inverse STFT ISTFT) to obtain a time-domain signal.
3. Apply STFT to the signal to obtain a new complex spectrogram.
4. Replace the magnitude with the original predicted magnitude while retaining the updated phase.
5. Repeat steps 24 for a fixed number of iterations (e.g., 50-100).

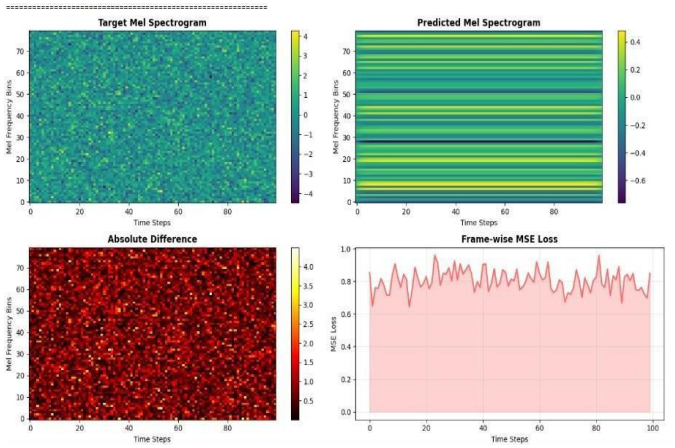

Figure 4: Visual representation of mel-spectogram

While as shown in the figure 4 Griffin–Lim is computationally heavier than a single-pass neural vocoder and may introduce artifacts, it has the advantages of:

- Simplicity and independence from additional training
- Reasonable quality when combined with high-quality mel predictions
- Full interpretability of the reconstruction process

In this work, Griffin–Lim is used as a practical, training-free vocoder, enabling end-to-end evaluation of the emotion-conditioned Tacotron model without the complexity of training a separate neural vocoder.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

6.1 Convergence and Spectrogram Quality

Training curves show that mel-spectrogram loss decreases from relatively high initial values to a stable low plateau. Visual inspection of predicted mel-spectrograms versus ground truth reveals:

- Accurate reproduction of formant structures and broad-band spectral patterns.
- Good temporal alignment of phonetic segments, indicating successful attention learning.
- Some residual discrepancies at high frequencies and sequence boundaries, which are common in Tacotron-style models, especially under purely autoregressive decoding.

Frame-wise error often increases toward the end of sequences, reflecting the cumulative effect of autoregressive prediction errors. This can be mitigated in future work with techniques such as scheduled sampling or non-autoregressive decoding.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 12, pp 192-198, December-2025

https://doi.org/10.47001/IRJIET/2025.912029

6.2 Emotion-Specific Synthesis

The model is evaluated qualitatively by synthesizing the same text under different emotion labels. The figure 5 shows Emotion-dependent characteristics observed in waveforms and spectrograms include:

- Happy / Surprise: Higher average pitch, greater pitch range, increased energy, and shorter segment durations. Waveforms show larger amplitude variation and denser high-frequency energy.
- Angry: Elevated energy and sharp, abrupt transitions. Spectrograms show concentrated high energy bands and more abrupt formant transitions.
- Sad: Lower pitch and reduced energy with elongated phoneme durations. Waveforms appear smoother with smaller amplitude envelopes; spectrograms show less high-frequency energy and slower temporal changes.
- Neutral: Baseline prosody with moderate pitch and energy. Spectral and temporal patterns fall between the extremes of the other emotion categories.

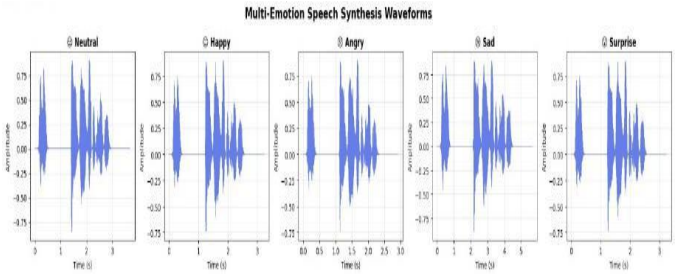

Figure 5: Multi-Emotion Speech Synthesis Waveforms

Listening tests (informal) indicate that human listeners can reliably distinguish between the emotional categories based solely on synthesized speech, suggesting that the model successfully maps emotion labels to characteristic prosodic patterns.

6.3 Web Application and User Interface

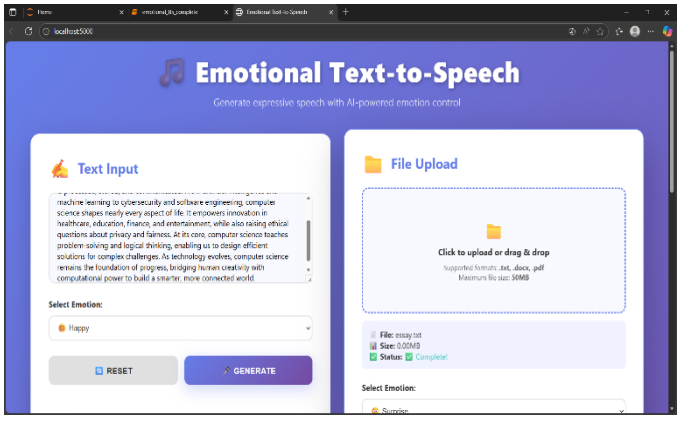

Figure 6: Text and File Input

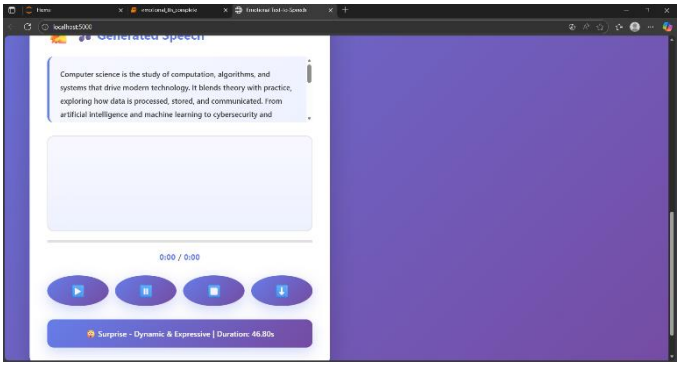

Figure 7: Generated Speech

The figure 6 and figure 7 represents web front-end built using Flask provides:

- Text input box for arbitrary sentences.
- File upload mechanism (e.g., .txt or .docx) for batch synthesis.
- Dropdown menu to select target emotion.
- Playback and download functionality for generated .wav files.

This demonstrates the practicality of the proposed system beyond theoretical research, offering a foundation for real-world deployments in applications requiring emotional TTS.

6.4 Technical Contributions

The main technical contributions of this work are:

1. Explicit Emotion Conditioning in Tacotron: Rather than relying on implicit style variation, the model treats emotion as a first-class conditioning variable, yielding controllable affect in synthesized speech.
2. End-to-End Pipeline from Emotional Speech Corpus: A complete workflow is established, from ESD-based preprocessing through neural training to waveform reconstruction and user-facing deployment.
3. Emotion-Aware Prosodic Patterns: Qualitative analysis confirms that learned prosodic patterns align with established SER findings on pitch, energy, and tempo variations across emotions.

6.5 Limitations

Despite promising results, several limitations remain:

- Griffin–Lim Vocoder: While simple, Griffin–Lim is computationally intensive and may produce artifacts, especially for long utterances. Neural vocoders such as WaveGlow, WaveRNN, or HiFi-GAN can produce higher fidelity and faster synthesis but require additional training and GPU resources.

ISSN (online): 2581-3048
Volume 9, Issue 12, pp 192-198, December-2025
https://doi.org/10.47001/IRJIET/2025.912029

- Limited Speaker and Language Diversity: This work focuses on the English portion of ESD and a restricted set of speakers, which may limit generalization to diverse accents and speaking styles.
- Lack of Quantitative Perceptual Evaluation: Objective metrics (loss curves, spectrogram comparisons) are informative but do not replace subjective listening tests such as Mean Opinion Scores MOS) and emotion recognition accuracy by human raters.
- Discrete Emotion Categories: Emotions are modeled as discrete labels, whereas real-world emotional states are continuous and context-dependent. This may limit the granularity of controllable expressiveness.

6.6 Future Work

Future directions include:

1. Neural Vocoder Integration: Replacing Griffin–Lim with a high-quality neural vocoder to improve naturalness and reduce artifacts.
2. Multi-Speaker and Cross-Lingual Training: Extending the architecture to multi-speaker scenarios and additional languages to enhance robustness and applicability.
3. Continuous Emotion Representations: Incorporating dimensional emotion models (e.g., valence–arousal) to provide more nuanced control over expressiveness.
4. Formal User Studies: Conducting systematic listening tests to measure perceived naturalness, intelligibility, and emotional clarity, as well as comparing against baseline neutral TTS systems.

VII. CONCLUSION

This paper presented an emotion-infused TTS system based on a Tacotron architecture extended with explicit emotion conditioning. Using the Emotional Speech Database and a carefully designed preprocessing and training pipeline, the model learned to generate mel spectrograms that, after Griffin–Lim reconstruction, produce speech with distinguishable emotional characteristics across five categories. Qualitative analysis of spectrograms, waveforms, and informal listening indicates that the system successfully enhances expressiveness while maintaining intelligibility.

The combination of explicit emotion embeddings, Tacotron-based sequence modeling, and a practical web interface demonstrates the feasibility of deploying emotionally aware TTS systems in real-world human–computer interaction scenarios. Although further work is needed to improve audio fidelity, generalization, and evaluation, the results confirm that deep learning–based emotion conditioning is a promising approach to bridging the gap between neutral synthetic speech and human-like expressive communication.

ACKNOWLEDGEMENT

It is agreat pleasure for us to present the project report on “Text to speech with emotion control using deep learning”, and we are thankful to all those who have helped us directly and indirectly for the successful completion of the project work.

We would like to express heartfelt gratitude to our respected guide, Mrs. Manasa SM, Assistant Professor, Department of Information Science and Engineering (IS&E), for her continuous encouragement and invaluable guidance throughout the project work.

We extend our thanks to our Project Coordinators for their unwavering support and encouragement during the course of the project.

We are also thankful to Dr. Raghavendra R. J., Associate Professor and Head, Department of IS&E, JNNCE, Shivamogga, and Dr. Y. Vijaya Kumar, Principal, JNNCE, Shivamogga, for their immense support and motivation.

We grateful to the Department of Information Science and Engineering and our institution, Jawaharlal Nehru New College of Engineering, for imparting us with the knowledge and resources to carry out our work to the best of our abilities.

Finally, we would like to thank the entire teaching and non-teaching staff of the Information Science and Engineering Department, as well as our families, for their constant support and encouragement.

REFERENCES

[1] X. Gao et al., “TTSlow: Slow Down Text-to-Speech with Efficiency Robustness Evaluations,” *IEEE Trans. Audio, Speech, and Language Process.*, 2025.

[2] L. Abdel-Hamid et al., “Analysis of Linguistic and Prosodic Features of Bilingual Arabic–English Speakers for Speech Emotion Recognition,” *IEEE Access*, vol. 8, pp. 7295772970, 2020.

[3] S. Seo et al., “Convolutional Neural Networks Using Log Mel-Spectrogram Separation for Audio Event Classification with Unknown Devices,” *J. Web Eng.*, vol. 21, no. 2, pp. 497522, 2022.

[4] O. Ghahabi and J. Hernando, “Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 25, no. 4, pp. 807817, 2017.

[5] Y. Masuyama et al., “Griffin–Lim Like Phase Recovery via Alternating Direction Method of Multipliers,” *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 184188, 2019.

ISSN (online): 2581-3048
Volume 9, Issue 12, pp 192-198, December-2025
https://doi.org/10.47001/IRJIET/2025.912029

[6] K. L. Ong et al., "Mel-MViTv2 Enhanced Speech Emotion Recognition with Mel Spectrogram and Improved Multiscale Vision Transformers," *IEEE Access*, vol. 11, pp. 108571 108579, 2023.

[7] R. Liu et al., "Modeling Prosodic Phrasing with Multi-Task Learning in Tacotron-Based TTS," *IEEE Signal Process. Lett.*, vol. 27, pp. 14701474, 2020.

[8] X. Tan et al., "NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 42344245, 2024.

[9] D. Yoshioka et al., "Nonparallel Spoken-Text-Style Transfer for Linguistic Expression Control in Speech Generation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 33, pp. 333346, 2025.

[10] A.Mukhamediya and A. Zollanvari, "On the Effect of Log-Mel Spectrogram Parameter Tuning for Deep Learning-Based Speech Emotion Recognition," *IEEE Access*, vol. 11, pp. 6195061957, 2023.

[11] D. B. de Souza et al., "Multitaper-Mel Spectrograms for Keyword Spotting," *IEEE Signal Process. Lett.*, vol. 29, pp. 20282032, 2022.

[12] R. Nenov et al., "Accelerated Griffin–Lim Algorithm: A Fast and Provably Converging Numerical Method for Phase Retrieval," *IEEE Trans. Signal Process.*, vol. 72, pp. 190202, 2024.

[13] R. Sato, R. Sasaki, N. Suga and T. Furukawa, "Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition," *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar*, 2020.

[14] Y. Zhang, "A Study on the Translation of Spoken English from Speech to Text," *in Journal of ICT Standardization*, vol. 12, no. 4, pp. 429-441, December 2024.

[15] Z. Liang, Z. Ma, C. Du, K. Yu and X. Chen, "E3TTS: End-to-End Text-Based Speech Editing TTS System and Its Applications," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4810-4821, 2024.

Citation of this Article:

Manasa S M, Subramanya S Gujjar, Shashank H S, Subhash G K, & Vikyath M A. (2025). Text To Speech with Emotion Control Using Deep Learning. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(12), 192-198. Article DOI https://doi.org/10.47001/IRJIET/2025.912029

****

