

Prosody Aware Speech Synthesis

Ms. Dali KU¹, Ms. Fathimath Farahath KA², Ms. Gayana KP³, Ms. Haritha C⁴, Mrs. Latha KV⁵
^{1,2,3,4,5} *Department of Computer Science and Engineering, Shree Devi Institute of Technology, Mangaluru, India*

Abstract—Standard text-to-speech (TTS) systems can generate understandable speech; however, they provide only limited control over the speaker identity, emotional expression, and detailed prosody. This paper introduces a modular, prosody-aware framework for multilingual expressive speech synthesis that implements language segmentation, transformer-based emotion analysis, continuous emotion-intensity estimation, adaptive prosody control, speaker-conditioned synthesis, and waveform-level audio processing. It is implemented as a centralized SpeechEngine plus a Streamlit interface. English and Hindi are the target languages for the speaker-conditioned voice cloning, and code-switched input was segmented into language spans before synthesis. The emotion analysis reports probabilities for six different target classes: happy, sad, angry, fear, calm, and neutral. A continuous value of intensity is extracted from the probability distribution using a confidence/entropy formulation and then mapped to conservative targets for pitch, speaking rate, energy, and pause duration. For the voice personalization, it was verified if a reference recording was provided, and XTTS-v2 was used if the required model and hardware capabilities were present. The resulting waveform was subjected to Librosa/SciPy-based digital signal processing, crossfading, peak-normalization, acoustic analysis, and explainability logging. The project artifacts attest to successful end-to-end synthesis and English/Hindi XTTS voice-cloning runs; however, the supplied validation documents also show the environment-dependent nature of XTTS availability and a missing complete human perceptual study. Therefore, the paper regards the reported system results as implementation-level evidence rather than proof of human-perceived superiority of emotion.

Index Terms—Text-to-Speech, Prosody Control, Expressive Speech Synthesis, Emotion Detection, Voice Cloning, Multilingual Speech, BERT, DistilBERT, XTTS-v2, Pitch, Speaking Rate, Energy, Pause Duration.

I. INTRODUCTION

Speech is one of the most natural modes of communication in humans, conveying not only verbal message but also emotional and emphasis content, speaking style, and identity of the talker. Conventional Text-to-Speech systems are mainly focused on turning text to speech in a manner understandable by the average listener. However, speech that lacks appropriate intonation, speaking rate, loudness, and pauses may sound dull or robotic.

This problem is especially evident in narration, educational talks, virtual assistants, accessibility systems, and personalized communication, where a dull or inappropriate expression may fail to deliver essential meaning and context. Even if a text-based message is linguistically correct, its robotic intonation fails to convey the emotional intent of the message sender.

Personalized speech synthesis poses another challenge in speaker-specific text-to-speech synthesis. Every talker has a set of characteristics peculiar to them, and a personalized assistant should be able to reproduce the traits of a particular person while retaining the contextual information and message.

Multilingual text and code-switching introduce additional difficulties, as a single sentence may have portions in multiple languages, with code-switching in both directions. The linguistic structure of such sentences may vary significantly, posing additional difficulties for automatic synthesis. The proposed project focuses on multilingual and code-switch text-to-speech synthesis for English and Hindi.

Recent advances in natural language processing and speech synthesis enable the combination of speaker-specific emotion recognition, multilingual processing, and explicit prosody control for the first time. This project aims to combine all these features into a single pipeline capable of producing expressive,

personalized, multilingual, and code-switch speech synthesis. The objectives of the project include:

- Natural and expressive speech production with control over emotion and prosody;
- Personalized speech synthesis with the use of reference speaker characteristics;
- Multilingual and code-switch speech synthesis for En-english and Hindi;
- Interpretable control over prosody in terms of pitch, speaking rate, loudness, and pauses.

II. LITERATURE REVIEW

Recent work in expressive and personalized speech synthesis has focused on improving the naturalness, emotional expressiveness, speaker identity, and multilingual capabilities of synthesized speech. Sang-Hoon Lee et al. [1] proposed a hierarchical speech representation approach that links semantic and acoustic representations for zero-shot speech synthesis. The work highlights the importance of hierarchical modeling in achieving high-quality speech synthesis with better speaker characteristics. Ha-Yeong Choi et al. [2] explored personalized and controllable voice style transfer revealing the importance of controlling speaker characteristics and speech style to achieve personalized expressive speech synthesis.

Prosody and emotional control have also attracted significant focus in recent speech synthesis research. Ismail Rasim Liao et al. [3] assessed objective evaluation of prosody and intelligibility in speech synthesis stressing the importance of assessing speech quality in terms of prosodic features in addition to intelligibility. Yi Zhou et al. [4] presented language-agnostic speaker embeddings for cross-lingual personalized speech synthesis. The work is pertinent to multilingual speech synthesis because it addresses the need for preserving speaker characteristics in cross-lingual speech synthesis. Kun Zhou et al. [5] studied emotion intensity and its control for emotional text-to-speech conversion highlighting the need for controlling the intensity of emotion rather than relying on discrete emotions in emotional speech synthesis. Several works have studied efficient speech synthesis, style control, and speaker cloning. Ali Raheem Mandeel et al. [6] investigated retraining-free pruning for text-to-speech synthesis models in speaker cloning. This approach seeks to reduce the complexity of text-

to-speech models for speaker cloning while preserving speech quality. Kurniawati Azizah et al. [7] studied transfer learning, style control, and speaker reconstruction for low-resource multi-speaker text-to-speech synthesis. This work demonstrates the relevance of transfer learning and style control in multilingual and multi-speaker speech synthesis. Kun Zhou et al. [8] analyzed speech synthesis with mixed emotions illustrating that emotional speech synthesis may require more than one emotional category.

Voice cloning and multilingual speech synthesis have also been explored in various works related to personalized text-to-speech synthesis. Yan Zhu et al. [9] presented a voice cloning method based on dual attention mechanisms and logistic distribution to enhance speaker identity and voice cloning quality in text-to-speech synthesis. Sathvik Udapa et al. [10] investigated multilingual Indian-language text-to-speech with voice cloning focusing on multilingual speech synthesis with preserved speaker identity. These studies illustrate the importance of integrating multilingual and voice cloning features in personalized text-to-speech synthesis.

While the above works address some of the main aspects of personalized text-to-speech synthesis such as emotion control, prosody, voice cloning, multilingual synthesis, and speaker embeddings, there is a need for an integrated pipeline that encompasses emotion detection, continuous emotion-intensity estimation, adaptive prosody mapping, personalized voice synthesis, audio crossfading, normalization, and acoustic analysis. The proposed Prosody Controlled Multilingual Expressive Speech Synthesis Using Voice Cloning approach addresses these aspects by developing a pipeline that integrates language segmentation, emotion detection and expression, adaptive prosody control, personalized voice cloning, and acoustic analysis. The system accepts English and Hindi input and translates the emotional content into prosody features such as pitch, speaking rate, energy, and pause duration to achieve expressive personalized speech synthesis.

III. METHODOLOGY

The proposed approach has eight steps: text input, multilingual language detection/segmentation, emotion detection, emotion intensity calculation, adaptive prosody mapping, personalized speech

synthesis, audio processing, and explainability/evaluation.

A. Text Input

The user provides the text that needs to be synthesized and can optionally provide a reference voice sample. The text can contain English, Hindi, and/or code-switched/positional multilingual text.

B. Multilingual Language Detection/Segmentation

The text is analyzed for the presence of English and Hindi segments, language segmentation is performed, code-switched text is handled, and the text segments are sent to the corresponding synthesis configuration.

C. Emotion Detection

An emotion detection model based on BERT/DistilBERT is used for computing the probability distribution over six emotions: Happy, Sad, Angry, Fear, Calm, and Neutral.

D. Emotion Intensity Calculation

The detected emotions are then converted into a continuous value between 0.00 and 1.00 to represent the intensity of the emotion. This allows the text to have varying degrees of emotional coloring rather than just a single emotion.

E. Adaptive Prosody Mapping

The emotion and intensity are converted into changes in pitch, speaking rate, energy, and pause duration for the generated speech. More intense emotions are associated with greater variations in these prosody features.

F. Personalized Speech Synthesis

The reference voice is fed into a speaker-conditioned TTS/voice-cloning model, which synthesizes speech with the given speaker characteristics while also incorporating the target language, emotion, and other prosody-related features. XTTS-v2 will be used if the execution environment permits.

G. Audio Processing

All language segments are synthesized individually and combined using equal-power crossfading before applying peak-normalization to generate the final WAV file.

IV. IMPLEMENTATION

The proposed Prosody Controlled Multilingual Expressive Speech Synthesis Using Voice Cloning is implemented as a modular Python-based speech synthesis framework integrated with a Streamlit application. The system is structured to unify language processing, emotion recognition, emotion-intensity estimation, prosody control, speaker conditioning, text-to-speech synthesis, digital signal processing, acoustic analysis and explainability into one unified processing pipeline.

A centralized SpeechEngine coordinates the major components of the system. Instead of having each application individually implementing text-to-speech synthesis, applications interact with a shared speech engine. This ensures that there is a consistent synthesis workflow across the various applications and to avoid superfluous duplication of model-loading and audio-processing operations.

A. System Initialization and Application Architecture
Upon launching, the Streamlit interface initializes the shared speech-generation engine. Resource caching is employed to maintain the same engine instance for all user interactions. This is crucial as transformer-based emotion models and neural speech-synthesis models consume large amounts of memory and take significant initialization time.

The initialization layer is responsible for loading the primary services that the system needs. These include language processing, emotion detection, prosody generation, text-to-speech routing, audio processing, voice management, persistence and analysis services. The overall software organization can be visualized as:

Streamlit Application → SpeechEngine → Processing Services → TTS Backend → Audio Processing → Output

The system additionally checks what computational device is available. If a compatible CUDA-enabled GPU is available, model inference can be performed using GPU acceleration; else, the system reverts to CPU execution.

Model preloading is employed to reduce the overhead of repeated initialization. Once the required models have been loaded, subsequent synthesis requests reuse the initialized components as opposed to loading the entire model again.

B. Text Input Processing

The synthesis process begins when the user inputs arbitrary text into the application. The input does not have to be limited to the pre-established example sentences. The text is sent to the central speech engine along with the selected or saved voice profile.

The speech engine receives parameters corresponding to language, emotion, intensity, voice profile and synthesis quality. Auto-detection is provided, which means that the default workflow does not require the user to explicitly specify every parameter.

The overall interface can be expressed as: $\text{SpeechEngine.generate}(T, V, L, E, I, Q)$ (1)

in which T denotes the input text, V denotes the voice profile,

L denotes the language information, E denotes the emotion,

I denotes the intensity and Q denotes the synthesis quality.

As a consequence, the system is able to handle the same input across different applications while maintaining a shared underlying synthesis architecture.

C. Language Detection and Segmentation

Before speech generation, the language-processing stage determines the language characteristics of the input text. The current implementation explicitly supports English and Hindi script spans.

Portions in Latin script are treated as English and those in Devanagari script are treated as Hindi. The system is able to maintain separate language spans as opposed to handling an entire multilingual sentence as one single unseparated string. An example of a code-switched input that includes English and Hindi portions may be expressed as:

$$T = [(t1, L1), (t2, L2), \dots, (tn, Ln)] \quad (2)$$

in which t_i denotes an individual text span and L_i denotes its detected language.

This segmentation allows for each span to be sent to the appropriate synthesis capability. It also provides the language information that the emotion and TTS stages need later.

D. Emotion Recognition Module

Prior speech synthesis, the emotion-recognition module analyses the textual input. For the supported emotion-classification path, the implementation leverages the Transformer checkpoint bhadrash-

savani/bert-base-uncased-emotion.

The model outputs a probability distribution across emotion

categories. These outputs are mapped to the project-level emotion set:

- Happy
- Sad
- Angry
- Fear
- Calm
- Neutral

Instead of retaining just the class with the highest probability, the complete probability distribution is retained. This allows for the later intensity-estimation stage to take into account both the dominant emotion as well as how uncertain the prediction is.

For a given input text T , the classifier outputs:

$$P(E|T) = [p1, p2, \dots, pK] \quad (3)$$

where p_i denotes the probability attached to emotion i .

The dominant emotion is determined as follows:

$$Edom = \arg \max_i p_i \quad (4)$$

The maximum probability corresponding to it becomes the confidence value for the following processing.

In the case of Hindi input, the current implementation also employs a lexical/heuristic mechanism. Consequently, the system does not consider the English Transformer checkpoint as a validated model for Hindi emotion recognition.

E. Emotion Probability and Confidence Processing

The probability distribution produced by the emotion model is utilized throughout the rest of the pipeline. This prevents emotion information from being calculated separately at multiple stages.

The confidence of the dominant emotion is defined as:

$$C = \max(P) \quad (5)$$

where P denotes the emotion probability vector.

An example case could have one emotion with a much higher probability value than the other classes, which would yield a higher confidence value. Alternatively, when the probabilities are more evenly distributed among the emotion categories, the confidence will be lower

The probability vector is also retained for display purposes.

F. Continuous Emotion Intensity Estimation

Rather than limiting emotion representation to just a categorical label, the implementation calculates a continuous emotion intensity value. The implementation combines the dominant probability with the entropy of the probability distribution.

The intensity is computed as follows:

$$I = P_{\text{top1}} \left(1 - \frac{H(P)}{\log_2 K} \right) \quad (6)$$

where P_{top1} denotes the highest predicted emotion probability, K represents the number of emotion categories, and $H(P)$ stands for the Shannon entropy.

The entropy is calculated as:

$$H(P) = - \sum_{i=1}^K p_i \log_2(p_i) \quad (7)$$

In the current implementation, $K=6$. The resultant value is bounded as:

$$0 \leq I \leq 1 \quad (8)$$

This value indicates the intensity of the detected emotional signal that the synthesis controller employs. It ought to be considered as a control parameter at the implementation level rather than a clinically or psychologically validated measure of human emotion.

F. Adaptive Prosody Controller

The Prosody Service translates the emotion information into control parameters at the speech level. This stage is in place to modify the acoustic properties of the generated speech according to the detected emotional state.

The system considers:

- 1) Modification of pitch/ F_0
- 2) Modification of speaking rate
- 3) Modification of energy
- 4) Modification of pause duration

The effective prosodic strength is calculated as follows:

$$S = I \times C \quad (9)$$

Here, I represents the emotion intensity while C denotes the emotion confidence.

The strength that results determines how strongly the corresponding preset prosody parameters for each particular emotion must be applied.

Since excessive post-processing can introduce unnatural artifacts, the implementation employs conservative modification ranges. The target configuration currently in use is as follows: The web application user interface part is created in Stream-lit. A single SpeechEngine coordinates the whole

synthesis process. It contains language processing, emotion detection, emotion intensity estimation, prosody generation, speaker conditioning, and text-to-speech components.

TABLE I Emotion Prosody Preset Configurations

Emotion	Pitch	Rate	Energy	Pause
Happy	+0.40 st	1.02×	1.03×	120 ms
Sad	-0.40 st	0.97×	0.97×	220 ms
Angry	+0.30 st	1.02×	1.03×	100 ms
Fear	+0.40 st	1.02×	1.02×	130 ms
Calm	-0.30 st	0.98×	0.98×	180 ms
Neutral	0.00 st	1.00×	1.00×	150 ms

V. CONCLUSION

The proposed system for Prosody Controlled Multilingual Expressive Speech Synthesis Using Voice Cloning provides an integrated approach to produce expressive, multilingual, personalized, and natural speech. The system combines multi-lingual segmentation, emotion detection, continuous intensity estimation, adaptive prosody control, speaker-conditioned synthesis, and acoustic analysis.

At the level of language-analysis, English and Hindi spans are detected in multilingual or code-switched input. The emotion component, which is based on BERT/DistilBERT, considers six emotions, and outputs probability information in a continuous intensity from 0.00 to 1.00. The calculated emotion-intensity is mapped onto pitch, speaking rate, energy, and pause duration.

The personalized speech is generated along with reference voice, whereas the separate language segments are merged using equal-power crossfading and peak normalization. The analysis of the waveform, pitch/ F_0 , and RMS energy provide interpretable acoustic information about the output. Thus, this project provides a unified pipeline, where the linguistic content, emotion, prosody, and speaker identity are simultaneously considered.

REFERENCES

- [1] S.-H. Lee, H.-Y. Choi, S.-B. Kim, and S.-W. Lee, "HierSpeech++: Bridging the Gap Between Semantic and Acoustic Representation of Speech by Hierarchical Variational Inference for Zero-

- Shot Speech Synthesis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 10, pp. 18422–18436, 2025, doi: 10.1109/TNNLS.2025.3584944.
- [2] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, “Personalized and Controllable Voice Style Transfer With Speech Diffusion Transformer,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 922–934, 2025, doi: 10.1109/TASLPRO.2025.3533362.
- [3] I. R. U’lgen, Z. Du, J. Lu, P. Koehn, and B. Sisman, “Objective Evaluation of Prosody and Intelligibility in Speech Synthesis via Conditional Prediction of Discrete Tokens,” *IEEE Open Journal of Signal Processing*, vol. 7, pp. 247–256, 2026, doi: 10.1109/OJSP.2026.3653666.
- [4] Y. Zhou, X. Tian, and H. Li, “Language Agnostic Speaker Embedding for Cross-Lingual Personalized Speech Generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3427–3439, 2021, doi: 10.1109/TASLP.2021.3125142.
- [5] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, “Emotion Intensity and Its Control for Emotional Voice Conversion,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2023, doi: 10.1109/TAFFC.2022.3175578.
- [6] A. R. Mandeel, T. Z. Fadhil, M. H. Alsalihi, M. S. Al-Radhi, et al., “Retraining-Free Pruning Text-to-Speech Synthesis Model for Speaker Cloning,” *IEEE Access*, vol. 13, pp. 193603–193615, 2025, doi: 10.1109/ACCESS.2025.3632214.
- [7] K. Azizah and W. Jatmiko, “Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages,” *IEEE Access*, vol. 10, pp. 5895–5911, 2022, doi: 10.1109/ACCESS.2022.3141200.
- [8] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, “Speech Synthesis With Mixed Emotions,” *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2023, doi: 10.1109/TAFFC.2022.3233324.
- [9] Y. Zhu, Y. Kang, and G. Chen, “A Voice Cloning Approach Utilizing Dual Attention Mechanism and Mixed Logistic Distribution,” *IEEE Access*, vol. 11, pp. 45120–45132, 2023.
- [10] S. Udupa et al., “Multilingual Indian-language text-to-speech with voice cloning,” in *Proc. IEEE Spoken Language Technology Workshop*, 2024, pp. 104–111.