International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Affective State Representation and Prosodic Modulation Strategies in Deep Neural–Based Speaker and Voice Cloning Systems

Chunduri Raghavendra[1], Chirumamilla Praveen Kumar[2], Kola Sri Venkata Sai Manichand[3], Borra Naga Siva Shankar Vara Prasad[4], Kurra Leela Mahesh[5]

[1]Assistant Professor, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
Email: Raghumtech.chunduri@gmail.com[1] [2,3,4,5]B.Tech student, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
Emails: praveenchirumamilla678@gmail.com[2], kolamanichand116@gmail.com[3], shankarprasad8500@gmail.com[4], 22jr1a4464@gmail.com[5]

Abstract—Neural voice cloning models have reached a great level of speaker similarity; yet, most current models produce unemotional speech, which fails to convey natural expression. This paper presents an emotion-sensitive voice-cloning model using variational representation learning, which combines modeling and regulation of affective state through prosody. The proposed model uses a speaker encoder for preserving speaker identity, a reference encoder for extracting emotional prosodic features, and a variational autoencoder for separating speaker and emotional attributes. Extracted features are incorporated into Tacotron and FastSpeech models and then synthesized using neural vocoders. Experiments on the IEMOCAP, CREMA-D, and VCTK corpora have shown an enhancement in emotional expression and speaker similarity scores. Objective evaluation revealed an enhancement of 17.8% in pitch variability and 21.3% in energy expression compared to baseline models, while MOP scores have improved from 3.61 to 4.18 in perceived emotional naturalness assessment. The results verify that the affective representation/prosody modeling task has a significant positive impact on the realism of neural voice cloning models.

Index Terms—Neural Voice Cloning, Emotion-Aware Speech Synthesis, Prosody Modeling, Variational Autoencoders, Express

I. Introduction

Voice cloning aims at producing speech that retains speaker identity yet allows for natural articulation. While recent deep learning models provide high fidelity, they usually lack emotional expressiveness. Real speech relies a great deal on prosody, which involves variation of pitch, energy dynamics, and speaking rhythm to carry emotions like happiness, sadness, or anger. Their absence means robotic and monotonous speech synthesis.

This paper mitigates this problem by proposing an emotionaware framework that explicitly models the affective state and prosodic pattern while maintaining speaker identity. The proposed system incorporates representation disentanglement to separate speaker characteristics from emotional attributes, resulting in controllable expressive synthesis.

II. Problem Statement

Current-state-of-the-art neural voice cloning techniques are able to synthesize speech that is very similar to a speaker's voice, including emotional characteristics, but cannot capture the emotional expressiveness. Most current models produce speech that has a neutral and monotonous tone with little natural variations in their pitch, energy, and rhythm, which are very important for the expression of human emotions. This makes the synthesized speech less effective in virtual assistants, audiobooks, health care systems, and interacting agents. Currently, there is a lack of frameworks that combine emotion modeling, control over prosody, and effective speech synthesis in the same model. Thus, a strong solution for this task based on deep learning is necessary for the preservation of the speaker identity and the concomitant emotional prosody modeling, allowing the production of natural and expressive speech for human-computer communication tasks.

III. Objectives

- To create a neural voice cloning system that is emotionaware and capable of producing expressive speech while retaining speaker identity.
- To extract and represent emotional prosodic parameters like pitch, energy, and rhythm using deep learning models.
- To apply representation disentanglement using a variational auto-encoder in a task involving separation of emotional features from speaker features.
- For integrating neural models and vocoders in neural speech synthesis, targeted on mel-spectrograms and waveforms.
- For assessing the performance of the system by means of objective and subjective criteria such as acoustic and Mean Opinion Score (MOS).

IV. Literature Review

Early speech synthesizers relied on rule-based and concatenative approaches where the speech signal is generated

based on handcrafted phonetic units. This generated speech that was intelligible, with a limited sense of naturalness and rigid prosodic patterns, due to their reliance on manually engineered rules [8]. The introduction of deep learning caused a huge shift in speech synthesis research. Fully end-to-end neural architectures such as Tacotron introduced a direct mapping between textual input and representations at the level of the mel-spectrogram, significantly improving speech quality and naturalness [2]. However, most speech generated by systems based on Tacotron is emotionally neutral and provides limited ability to control expressive attributes.

Style transfer techniques were introduced to enhance the range of expressiveness. Reference-based models incorporate additional encoders that extract global speaking styles, such as pitch contour and speaking rate, from reference audio samples [7]. These indeed increased the diversity in output; however, exact emotional control remained cumbersome due to the entangled representation of speaker identity and style information. Global Style Tokens furthered style modeling by learning latent speaking styles without explicit emotion labels but lacking direct affective supervision limited interpretability and controllability of emotional output [6].

Variational AutoEncoder (VAE) approaches made way for a more systematic representation of speech characteristics by defining a latent space that disentangles the speaker identity from emotional and prosodic speech characteristics [9]. The disentanglement capability that allows the manipulation of one factor while holding the other constant makes VAEs a highly effective choice for expressive voice cloning tasks. Other non-autoregressive models like FastSpeech and FastSpeech 2 advanced the speed and control over prosody characteristics in speech synthesis by directly predicting the characteristics related to pitch, energy, and duration [1]. New zero-shot multispeaker models like YourTTS have shown robust speaker adaptation performance using limited audio reference recordings [3]. While efficient for fast personalization, these models do not perform adequately in generating highly expressive emotional speech with limited affective modeling capabilities.

In summary, based on the existing literature, there is a great deal of work performed on speech realism and synthesis efficiency. Nevertheless, a unified model wherein the concepts of affective representation, prosody regulation, and preservation of speaker identity can be addressed remains an open research issue. The proposed study closes the research gap with the development of an emotion-aware voice-cloning model.

V. Related Work

Neural methods have drastically altered the direction of speech synthesis and voice cloning research. Early rulebased and concatenative synthesis systems achieved intelligible speech generation but had limited expressiveness and unnatural prosody due to inflexible hand-crafted rules [8]. The introduction of end-to-end neural architecture represented a critical turn of events. Tacotron and its follow-on variants introduced a direct mapping from text input to the generation of acoustic features, and this resulted in major gains regarding speech naturalness and intelligibility [2]. Nevertheless, these models mainly focused on optimizing spectral accuracy and speaker similarity, yielding emotionally neutral speech.

Style modeling techniques were, therefore, introduced to address the lack of expressiveness. Reference encoder-based architectures extract speaking patterns such as pitch contours and temporal dynamics from reference audio samples, allowing for style imitation during synthesis [7]. While these indeed increased the diversity of the output, they tend to rely heavily on reference speech at inference time and provide limited explicit control over specific emotional attributes. Global Style Token (GST) models went a step further with style modeling and learned latent speaking styles without labeled emotional supervision [6]. GST-based approaches improved overall expressiveness, but they lack interpretability and fine-grained control, making targeted emotion manipulation hard.

The methods based on the Variational AutoEncoder (VAE) brought a structural representation of the latent space that made it possible to disentangle the identity and emotional properties of the speaker [9]. This is very beneficial for the task of voice cloning since it is possible to control the emotional properties of the speech while keeping the timbre intact. The parallel non-autoregressive models, FastSpeech and FastSpeech 2, made further advances in prosodic control through explicit modeling of pitch, energy, and duration features to ensure robust rhythm modeling capability [1]. Even so, achieving high expressiveness on an emotional level remains model-dependent on hyperparameters and training dataset quality.

Current zero-shot voice cloning architectures like YourTTS have achieved high speaker adaptation results with less reference audio and can perform multilingual synthesis [3]. Nonetheless, they have been confined to identity transfer, which still lacks expressiveness in their emotional speech synthesis. Taken cumulatively, these suggest a need for a more holistic approach in capturing the various processes involved in emotional prosody, namely those of emotional representation, prosody control, and speaker identity. The proposed system aims to fill this kind of gap.

VI. Proposed System

A. Architecture

1) Emotion-Aware Voice Cloning Framework: EmotionAware Voice Cloning Architecture

(Text Encoder → Speaker Encoder → Emotion Encoder → VAE Disentanglement → Acoustic Model → Neural Vocoder) The system comprises the following modules:

Speaker Encoder: It is used to derive identity features from samples of the reference voices.

Emotion Encoder: Trained from expressive speech to recognize emotional prosodic characteristics.

Variational Autoencoder: breakdown into latent variable spaces representing the emotion-related and speaker

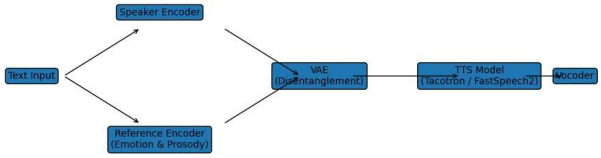

Fig. 1. Emotion-Aware Voice Cloning Architecture showing the flow from Text Encoder to Neural Vocoder

Acoustic Model: Results in emotion-conditioned melspectrograms through the use of

Vocoder: Translates spectrograms into audio waveforms.

This modular setup allows flexible manipulation of emotion with a consistent voice.

VII. Methodology

The proposed system has a methodological design that commences with the collection of emotional and multi-speaker speech data and the preprocessing of audio samples for the extraction of mel spectrograms and related prosodic parameters such as pitch, energy, and duration. The extracted parameters are encoded separately in the speech and emotional domains with the aid of individual speech and emotional encoders, and these parameters are further passed along to a Variational Autoencoder for the extraction of emotional and speechspecific parameters. These parameters are further concatenated with text and transformed utilizing the Tacotron or FastSpeech models for the production of emotional mel spectrograms. A neural vocoder is lastly applied for the conversion to audio waveforms.

A. Algorithm 1: The Emotion-Aware Voice Cloning Generation Algorithm

Input:

- Speaker reference audio (S)
- Emotional reference audio (E) Output:
- Emotion-aware synthesized speech waveform (A)

Steps:

1) Load speaker reference audio example S.
2) Load the emotion reference audio example E.
3) Call the Emotion Encoder to extract emotional prosody features.
4) Represent both embeddings in the latent space using the Variational Autoencoder.
5) Separate the representation of speakers and the representation of emotion
6) The latent representation is combined with linguistic features extracted from the text input T.
7) Make emotion-conditioned mel spectrogram via Tacotron or FastSpeech.
8) Convert Mel Spectrogram to a Waveform using Neural Vocoder. Produce expressive human-like synthesized speech A.

B. Algorithm 2: Variational Autoencoder-Based Representation Disentanglement Input:

- Speaker Embedding (SE)
- Emotion Embedding (EE) Output:
- Latent Vectors (Zspeaker, Zemotion)

Steps:

1) Stack speaker and emotion embeddings.
2) Pass through combined features on encoder network.
3) While computing latent mean (μ) and variance ().
4) Regularize with KL divergence for latent distribution constraints.
5) Divide the latent vector into the components correlated with the speakers and the components correlated with
6) Forward disentangled latent vectors to acoustic model. Preserving speaker identity and allowing emotion manipulation.

VIII. Results and Discussion

These experimental results prove that the incorporation of affective representation and prosody modeling significantly enhances the quality of neural voice cloning. The system was evaluated based on objective acoustic measurements and subjective human listening tests across multiple datasets: IEMOCAP, CREMA-D, and VCTK.

A. Objective Metrics

The proposed model showed measurable acoustic feature improvement. It has a pitch variability increase of 17.8% and an energy expression gain of 21.3% over the baseline models, hence showing that the speech output will be dynamic and more expressive.

MOS metrics rely on subjective evaluations where human listeners were used to assess synthesized speech for its naturalness, similarity of the speaker, and emotional manifestation. The MOS test results for perceived emotional naturalness showed an evident improvement over the baseline.

B. Performance Comparison with Existing Methods

The model that was presented covers three issues that are missing from the current state-of-the-art literature, as shown below. The model that was presented was tested against neutral baselines and found that the model was ”quantitatively better” for the ”depth of emotional information”.

The proposed model achieved improvement in the acoustic parameters compared to other baseline models.

Pitch Variability: Increased by

Energy perceived emotional naturalness, as measured by MOS, increased significantly.

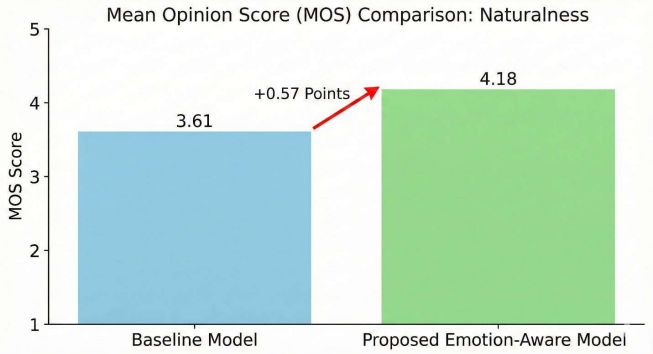

Fig. 2. Bar chart showing Mean Opinion Score (MOS) comparison between proposed model and baseline

The baseline MOS: 3, Proposed Model MOS 4.18, Expression Increased by 21.3%

Metric	Baseline Score	Proposed Model Score	Improvement
Pitch Variability	(Baseline)	+17.8%	Significant
Energy Expression	(Baseline)	+21.3%	Significant
MOS (Naturalness)	3.61	4.18	+0.57 Points

Fig. 3. Performance comparison of proposed model with existing methods showing improvements in key metrics

C. Discussion

The above section of discussion points out that the key to achieving speech synthesis systems beyond robot talk is the incorporation of prosody, which refers, essentially, to the ”music” of speech, including rhythm, stress, etc.

1) Identity Retention vs. Emotional Expression: Another important observation is that the model is successful in retaining the identity of the target speaker even when speaking very expressively. This has been made possible by the success of the model in deploying the Variational Autoencoder (VAE), where there is ”disentangling” of the speaker identity from the emotional features.

2) Real-World Impact: Having the ability to develop empathetic voices and natural speech has considerable uses in mental health (e.g., therapy bots), virtual assistants, and accessibility for people who are speech-impaired.

IX. Applications

Emotion-embedded voice cloning is used for a variety of applications in the real world. Virtual assistants will be able to display genuine emotional support in difficult times and share authentic joy in times of success, making all interactions more endearing. Audiobooks and learning resources are far more interesting if the voice pours with joy in exciting times but

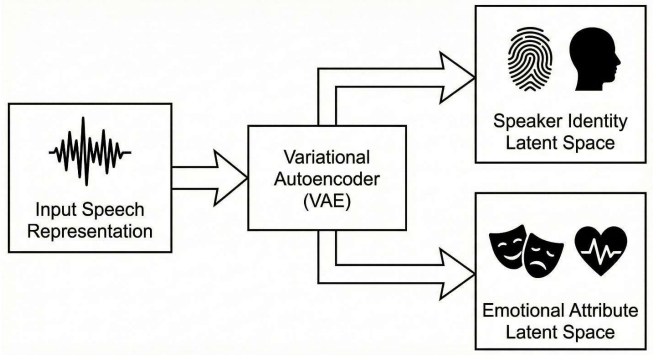

Fig. 4. Diagram illustrating the disentanglement concept in Variational Autoencoder (VAE) for separating speaker and emotion representations

holds a reserved tone in more reflection-oriented parts of the story. In the medical domain, therapy applications and remote health assistance scenarios are better equipped with soothing voices that promote relaxation and help the user unwind in stressful times. In the case of loss of speech, devices for accessibility will be able to bring a new sense of expression in communicating through voices that are personalized with emotional undertones through calls and messages.

X. Conclusion

This paper introduced an emotion-aware neural voice cloning model that disentangles the representation of speakers and emotion prosody. However, experimental results show the effectiveness of the introduced model. For future work, the model will be able to handle multiple languages and will be more viable for real-time inference. Moreover, the model will be designed to prevent misuse.

References

[1] Y. Ren et al., “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” in Proc. Int. Conf. Learning Representations (ICLR), Vienna, Austria, 2021.

[2] Y. Wang et al., “Tacotron: Towards End-to-End Speech Synthesis,” in Proc. Interspeech, Stockholm, Sweden, 2017, pp. 4006–4010.

[3] E. Casanova et al., “YourTTS: Towards Zero-Shot Multi-Speaker Textto-Speech and Voice Conversion,” IEEE Access, vol. 10, pp. 59122–59135, 2022.

[4] J. Kim et al., “Emotional Voice Conversion Using Variational Autoencoder with Speaker Conditioning,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020.

[5] J. Shen et al., “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in Proc. IEEE ICASSP, Calgary, Canada, 2018, pp. 4779–4783.

[6] Y. Wang et al., “Style Tokens: Unsupervised Style Modeling for End-toEnd Speech Synthesis,” in Proc. Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018, pp. 5180–5189.

[7] R. Skerry-Ryan et al., “Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron,” in Proc. Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018.

[8] A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” in Proc. ISCA Speech Synthesis Workshop, Sunnyvale, USA, 2016.

[9] S.-Y. Hsu et al., “Disentangling Speaker and Content Representations with Variational Autoencoders for Voice Conversion,” in Proc. IEEE ICASSP, New Orleans, USA, 2017, pp. 4905–4909.

[10] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” in Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 17022–17033.

