Beyond Mimicry: Emotion and Accent-Aware Voice Cloning

Aayushi Bhimani
IIT Jodhpur
m23cse001@iitj.ac.in

Prabhat Ranjan
IIT Jodhpur
m23csa017@iitj.ac.in

Abstract

Text-to-speech (TTS) systems have improved significantly in generating natural-sounding voices, but still struggle to express different emotions and diverse accents. In this work, we built an adaptive voice cloning system - 'EchoSpeech', that can not only copy a speaker's voice but also change the emotional tone and accent of the synthesized speech. We leverage Tortoise-TTS for high-quality voice cloning and added emotion control using reference samples from the CREMA-D dataset. For accent transfer, we fine-tuned a speaker embedding (SE) model based on Wav2Vec2, training it to recognize and represent different accents, allowing us to apply accent styles to cloned voices. The system can now produce speech that sounds not only like the original speaker but also happy, sad, angry, or fearful, and speak in various global accents such as Nigerian, Filipino, or Scottish. Our results show that the cloned voices remain clear and consistent while reflecting both the desired emotion and accent. This approach brings us closer to more expressive, personalized, and culturally adaptive TTS applications for education, entertainment, accessibility, and more.

1. Introduction

Text-to-speech (TTS) systems [3] have made remarkable progress in producing fluent and intelligible speech. However, most existing models are limited in their ability to express emotion or adapt to diverse accents while preserving the identity of the original speaker. These limitations reduce their effectiveness in scenarios where expressiveness and cultural adaptability are crucial, such as virtual assistants, storytelling, accessibility tools, and language education platforms.

In this work, we introduce a unified and modular pipeline that enables adaptive voice cloning with both emotion and accent control. Unlike traditional systems that handle these dimensions independently or not at all, our approach integrates emotional expressiveness and regional accent transformation into a single end-to-end framework. The system accepts a short reference audio clip and a text prompt, and produces speech that maintains the speaker's identity while adapting its tone and accent to user preferences.

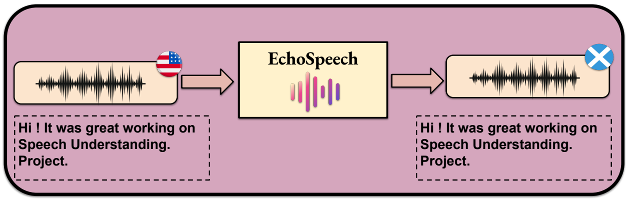
EchoSpeech
Hi ! It was great working on Speech Understanding. Project.
Hi ! It was great working on Speech Understanding. Project.

Figure 1. The central goal of this work is to synthesize speech that not only preserves the speaker's identity but also adapts expressively to different emotions and accents, using a unified pipeline combining voice cloning, emotional conditioning, and accent transfer.

The core of the system leverages Tortoise-TTS, a high-fidelity autoregressive model for voice cloning [3]. To enable emotion conditioning, we integrate reference samples from the CREMA-D dataset [4], guiding the synthesis process to produce speech with distinct emotional tones—neutral, happy, sad, angry, or fearful.

For accent control, we fine-tune a Wav2Vec2-based speaker embedding extractor on a labeled accent dataset [2]. This allows the model to capture accent-specific characteristics that are later used with OpenVoice's ToneColorConverter [5] to apply accent transformations while preserving vocal identity. The fine-tuning step enhances the extractor's sensitivity to accent cues, improving the realism of the transferred speech.

To ensure accessibility and ease of use, we deploy the entire system through a Gradio-based web interface [1]. This enables real-time, user-driven speech synthesis by allowing users to input custom text, upload reference audio, and select desired emotional and accentual characteristics.

Summary of Contributions

- We present a unified pipeline for expressive voice cloning that simultaneously supports emotion and accent control within a single framework.
- We adapt Tortoise-TTS [3] to synthesize emotion-aware speech using curated reference clips from the CREMA-D

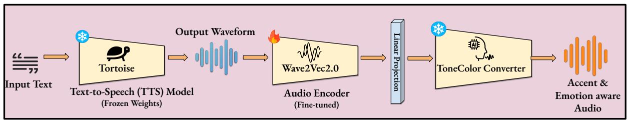

Figure 2. Proposed pipeline for expressive voice cloning with emotion and accent control. The system integrates Tortoise-TTS for voice cloning, emotional conditioning using CREMA-D references, Wav2Vec2-based accent embedding extraction, and accent transfer via OpenVoice's ToneColorConverter [5].

dataset [4].

- We fine-tune a Wav2Vec2-based speaker embedding model [2] for accurate accent representation and transfer.
- We integrate OpenVoice's ToneColorConverter [5] for realistic accent transformation while preserving speaker identity.
- We develop a user-friendly Gradio interface [1] that enables interactive and customizable speech synthesis through simple inputs.

This system offers a practical and extensible solution for expressive and culturally adaptive speech synthesis that preserves speaker identity while enhancing vocal style.

2. Methodology

Our proposed pipeline is designed to synthesize expressive speech that maintains the identity of the reference speaker while adapting to user-specified emotional tone and accent. The system is composed of four main stages: (1) voice cloning using Tortoise-TTS [3], (2) emotion reference conditioning [4], (3) accent embedding extraction and transfer [2], and (4) final speech output generation through an integrated interface [1]. An overview of the full pipeline is shown in Figure 2.

2.1. Voice Cloning with Tortoise-TTS

We employ Tortoise-TTS as the backbone of our voice cloning module due to its high-quality output and support for reference-based synthesis [3]. A short reference audio clip from the target speaker is provided, and the model generates speech that closely mimics the speaker's identity while reading the input text. This forms the base layer of our synthesis pipeline.

2.2. Emotion Conditioning

To infuse emotion into the generated speech, we use reference samples from the CREMA-D dataset [4]. We group emotion-labeled samples and create folders containing multiple reference clips (up to 5 per emotion) for each speaker-emotion combination. These clips are fed into Tortoise-TTS during inference to influence prosody and tone, enabling synthesis of emotionally expressive speech. The supported emotions are: *neutral*, *happy*, *angry*, *sad*, and *fearful*.

2.3. Accent Embedding Extraction

For accent control, we design and train a custom speaker embedding extractor based on the Wav2Vec2 model [2]. The extractor is fine-tuned using a labeled accent dataset consisting of speech clips from diverse regional accents. A projection head is added on top of the Wav2Vec2 backbone to map features to a lower-dimensional embedding space. The resulting speaker embeddings capture accent-specific characteristics and can be averaged across clips to form robust accent prototypes.

2.4. Accent Transfer with ToneColorConverter

We use OpenVoice's ToneColorConverter module to perform accent transfer [5]. Given a generated speech waveform and an accent embedding (from the extractor), the module modifies the speech to reflect the accent's tone and articulation while maintaining the original speaker's voice. This process enables high-quality, realistic accent adaptation across multiple target accents.

2.5. User Interaction via Gradio Interface

To make the entire pipeline accessible to non-technical users, we integrate the system into a web interface using Gradio [1]. Users can upload a reference audio clip, enter desired text, choose an emotion, and select an accent. The backend automatically processes the inputs, generates intermediate embeddings, and produces the final speech output—all within an interactive UI.

3. Experiments and Results

We conducted a series of qualitative experiments to evaluate the effectiveness of our proposed pipeline in generating expressive speech with controlled emotion and accent, while preserving speaker identity. The evaluation focuses on two aspects: (1) the quality of accent representations before and after fine-tuning the speaker embedding model, and (2) the variation in prosodic characteristics across emotion-conditioned speech outputs.

3.1. Accent Embedding Visualization

To assess the impact of fine-tuning on accent representation, we visualized speaker embeddings using t-SNE. Figure 3 displays embeddings generated by the initial, non-finetuned Wav2Vec2-based model. The accent classes are largely entangled, showing little structure or separation.

After fine-tuning the extractor on labeled accent data, the updated embeddings—shown in Figure 4—exhibit well-separated and dense clusters for each accent. This indicates that the model has learned discriminative, accent-specific features, which are critical for effective downstream accent transfer. In particular, the clusters for Nigerian, Filipino, and Scottish English become highly distinguishable.

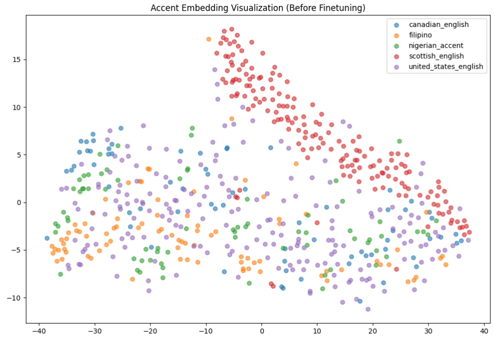

Figure 3. Accent embeddings before fine-tuning: Significant overlap among accent classes and poor separation in the latent space.

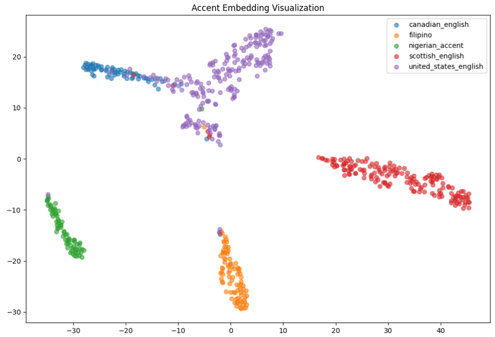

Figure 4. Accent embeddings after fine-tuning the Wav2Vec2-based extractor: Improved clustering and clearer separation across accent classes.

3.2. Emotion-aware Voice Synthesis

We further evaluated the ability of the system to synthesize emotion-conditioned speech using Tortoise-TTS guided by emotion reference clips. Mel spectrograms of generated outputs for five emotion categories—*neutral*, *angry*, *happy*, *sad*, and *fear*—are shown in Figure 5.

Notable differences in pitch, energy, and duration can be observed across the spectrograms. For instance, angry and happy samples exhibit higher energy and denser frequency content, while sad and fearful samples show more pauses and reduced spectral energy. These variations align with human prosodic patterns for emotional speech, confirming that emotion control is successfully embedded during synthesis.

3.3. Speaker Identity Preservation

To evaluate how well the generated speech retained the speaker's identity across different emotions and accents, we computed cosine similarity between speaker embeddings of the reference audio and the synthesized output. Table 1 presents these similarity scores.

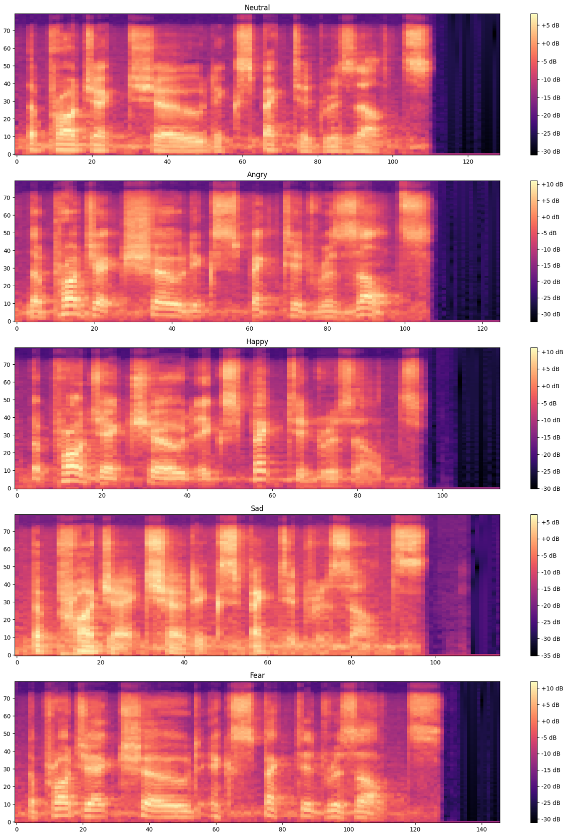

Figure 5. Mel spectrograms of synthesized speech conditioned on five emotions. Distinct differences in prosody and energy reflect successful emotion injection while preserving speaker identity.

Table 1. Cosine similarity between reference and generated voices across different emotions and accents. Higher values indicate better preservation of speaker identity.

Accent	Neutral	Happy	Angry	Sad	Fear
Canadian English	0.9525	0.8683	0.9463	0.9204	0.9178
Filipino	0.8630	0.8421	0.8511	0.9809	0.9242
Nigerian Accent	0.6037	0.7856	0.6072	0.7836	0.9233
Scottish English	0.9260	0.9678	0.9489	0.9381	0.7987
United States English	0.9800	0.5010	0.4493	0.9584	0.8397

The results show that identity preservation was strongest for neutral and sad emotional states, with similarity scores exceeding 0.95 for several accents. Lower scores were observed for high-variability conditions, such as angry or fearful speech in the United States English and Nigerian accents. This suggests that while emotional and accent modulation introduces variation, the core speaker characteristics remain largely intact.

3.4. Qualitative Observations

We conducted informal perceptual evaluations through the Gradio interface to assess the naturalness and expressiveness of the generated speech. Based on our observations, both emotional tone and accent characteristics were clearly reflected in the outputs. The synthesized speech was found to be intelligible, natural, and free of major artifacts. Accent transfer preserved the reference speaker's vocal identity, while emotion variations were distinct yet not overly exaggerated. These findings qualitatively support the effectiveness of the proposed pipeline for expressive and adaptive speech synthesis.

4. Conclusion and Future Work

This work presents a modular and interactive pipeline for adaptive voice cloning that incorporates both emotion control and accent transfer. Built upon the Tortoise-TTS backbone, our system utilizes curated emotion reference clips and a fine-tuned Wav2Vec2-based speaker embedding model to generate expressive speech that retains speaker identity across diverse emotional and accentual contexts. Integration with OpenVoice's tone color conversion and a Gradio interface enables seamless user interaction and real-time voice customization.

Qualitative results demonstrate that our system produces speech with clear emotional variation and perceptible accent characteristics, while maintaining high intelligibility and naturalness. Accent embeddings became more discriminative post fine-tuning, and spectrogram analyses verified successful emotion conditioning. Cosine similarity scores further validated the model's ability to preserve identity even under expressive transformations.

Future work will focus on expanding the system's capabilities by:

- Incorporating a broader range of emotions and multilingual accent types.
- Conducting formal perceptual studies with human raters for quantitative validation.
- Exploring unsupervised or few-shot learning methods for better generalization to unseen voices.
- Enhancing run-time performance for real-time deployment on edge devices.

Overall, the proposed pipeline provides a flexible and extensible framework for expressive and personalized text-to-speech synthesis, enabling applications in virtual assistants, entertainment, accessibility, and more.

References

[1] A. Abid, D. Abdul, J. Zhang, and et al. *Gradio: Build and share delightful machine learning apps*, 2021. https://gradio.app. 1, 2

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12449–12460, 2020. 1, 2

[3] J. Betker. Tortoise tts: A multi-voice text-to-speech system trained on diverse audio datasets. *GitHub Repository*, 2022. https://github.com/neonbjb/tortoise-tts. 1, 2

[4] H. Cao, E. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. 1, 2

[5] MyShell. Openvoice: Voice cloning with tone and style transfer, 2023. https://github.com/myshell-ai/OpenVoice. 1, 2