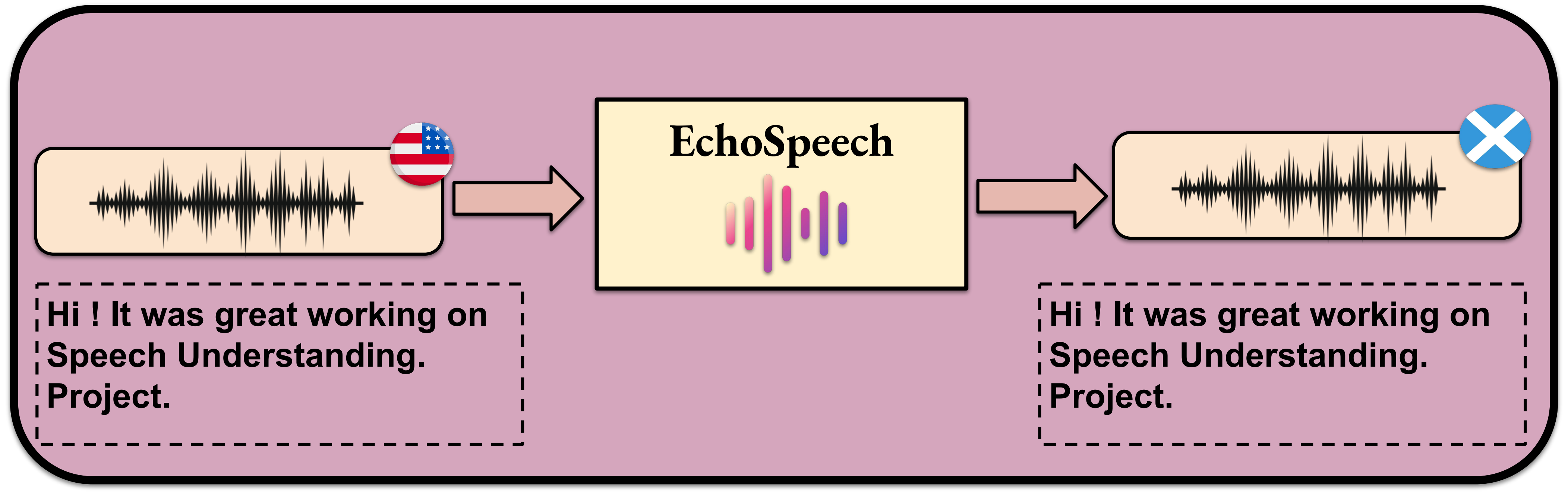
Beyond Mimicry: Emotion and Accent-Aware Voice Cloning

Aayushi Bhimani(M23CSE001), Prabhat Ranjan(M23CSA017)

IIT Jodhpur

Aim



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

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

Enhancing *cloned*, *emotion-rich* audio signals with "*accents*" using transformers.

Limitations of Existing Systems

- **Lack of emotional expressiveness: Most TTS systems generate speech with a flat tone and cannot convey emotions**
- **Existing TTS systems generally do not support dynamic accent adaptation**

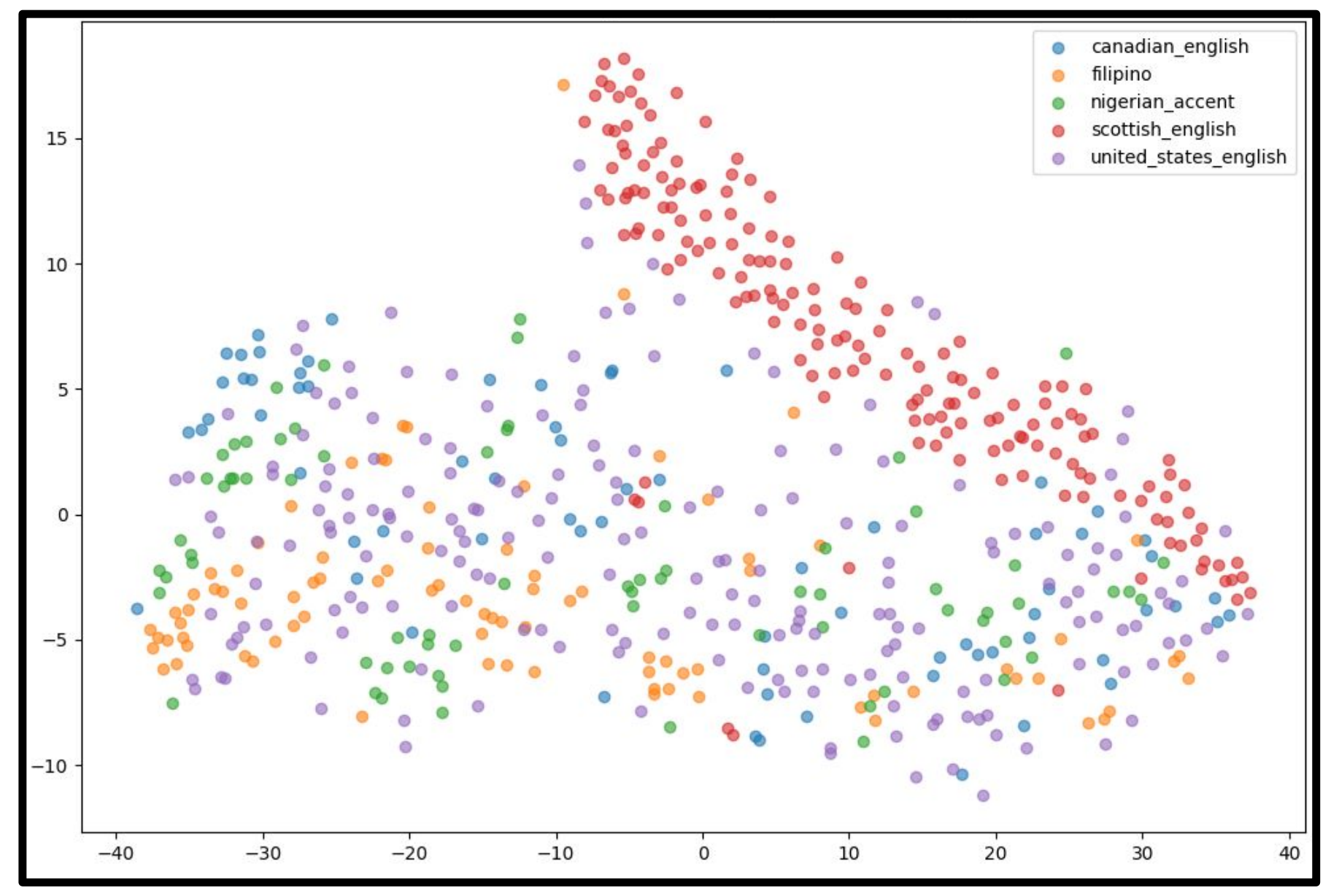
Experiments

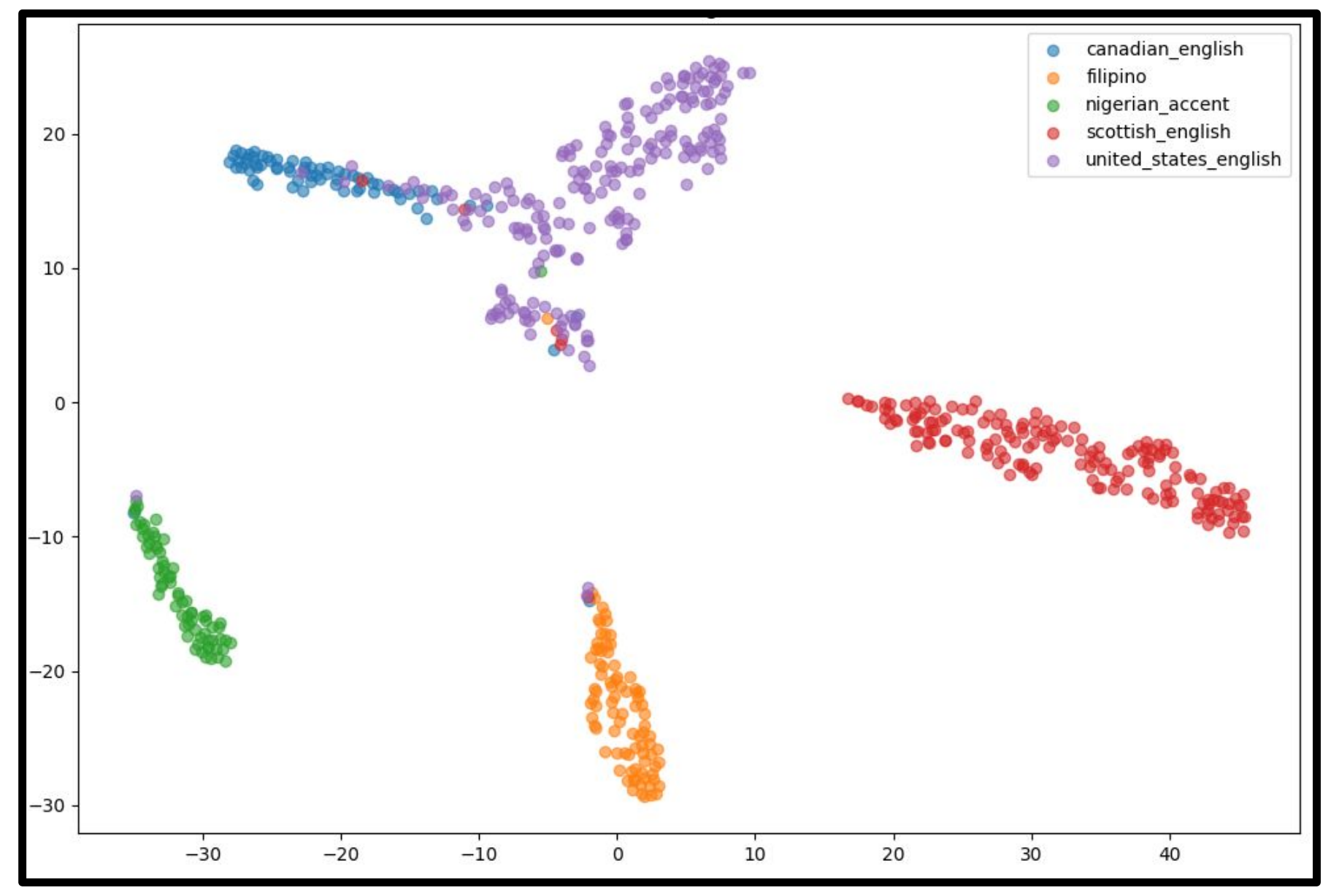
Table 2. Cosine similarity between reference and generated voices across emotions and accents. Higher values indicate better preservation of speaker identity.

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

- Cosine similarity scores show that a speaker's voice stays more consistent in neutral and sad emotions across most accents.
- In contrast, angry and happy emotions especially with accents like Nigerian or American English change the speaker's voice more and make it harder to recognize.

Qualitative Analysis

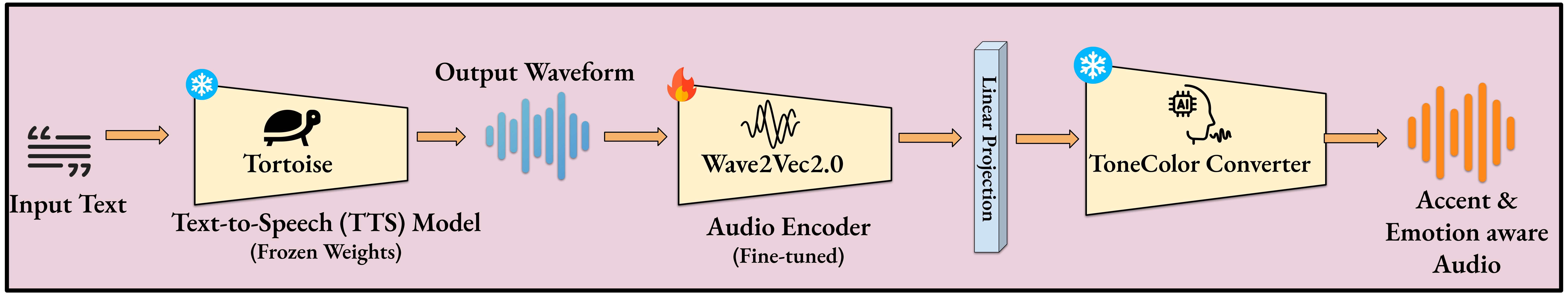




Before Fine-Tuning: Accent embeddings show substantial overlap, reflecting poor discriminability across accents.

After Fine-Tuning: Embeddings form distinct, well-separated clusters, indicating improved accent-specific representation.

Methodology



Conclusion

- **We present "*EchoSpeech",* a framework that incorporates emotion control using Tortoise-TTS, guided by curated reference audio clips from the CREMA-D dataset.**
- **We fine-tune a speaker embedding model based on Wav2Vec2.0 to learn discriminative accent features for effective accent classification.**
- **We integrate OpenVoice's Tone Color Converter to perform realistic accent transfer using learned speaker embeddings.**

References

1. J. Betker, *Tortoise-TTS: A Text-to-Speech System with Emphasis on Expressive Style and Identity Preservation*, arXiv:2305.07243, 2023.
2. R. Xu, C. Xu, and Y. Wu, *OpenVoice: A Versatile Instant Voice Cloning Approach with Emotion and Accent Control*, arXiv:2312.01479, 2023.