WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

ACCENOVA CONVERTER

Prajwal P Doddamani[a], Narayana Reddy N R[b], Mohammed Ayan[c], Manjunath B[d], Mr Vekateshwara N [e]

[a,b,c,d] Department Of Computer Science and Engineering, Jyothy Institute Of Technology, Bengaluru, India

[e] Guide, Department. Of Computer Science and Engineering, Jyothy Institute Of Technology, Bengaluru, India

A B S T R A C T :

English is spoken in many accents around the world, and although the language is the same, the way it sounds can change dramatically depending on the region. These accent differences affect pronunciation, rhythm, stress, and style, often resulting in misunderstandings during online communication, classrooms, interviews, and global teamwork. This project presents Accent Converter, a web-based system that modifies a person's accent while keeping their original voice. The system uses modern speech-processing techniques such as speaker encoders, prosody modeling, and neural vocoders. With only a few seconds of user speech, the system extracts voice identity, applies target accent patterns, and produces natural-sounding speech with the chosen accent. This paper discusses the motivation, system design, technical methodology, implementation pipeline, performance evaluation, limitations, and the future potential of the system. Results show clear changes in accent patterns while preserving the speaker's voice quality, demonstrating the usefulness of accent transformation models.

Keywords: Accent Conversion, Speech Processing, Prosody Transfer, Neural Vocoder, Voice Identity, Deep Learning, Speech Synthesis

1. Introduction

English is spoken globally, but the way it sounds varies widely across countries. Indian, American, British, and Australian speakers pronounce the same words differently due to variations in vowel shaping, consonant articulation, pitch flow, and timing. These differences collectively form what we call an "accent." While accents are natural and reflect cultural identity, they sometimes cause misunderstandings—especially when speakers are unfamiliar with each other's accent.

With the growth of online learning, international meetings, remote interviews, and content creation, the need for clearer and more neutral accents has grown. Many people also want to hear how they would sound in a different accent, either for education or personal interest.Most speech technologies today rely on text-to-speech (TTS). These can generate audio in any accent but replace the speaker's voice completely, making the output sound artificial or unfamiliar.

Modern deep-learning technologies, however, allow us to separate content, voice identity, and speaking style, meaning we can change how someone sounds without changing who is speaking. This project focuses on building Accent Converter, a web tool that allows users to convert their accent into another accent while keeping their voice.

Instead of training huge models from scratch, which requires enormous datasets, the project uses pre-trained systems and modifies them for practical usability.

2. Literature Survey

The use of artificial intelligence in accent conversion has advanced rapidly in recent years. Earlier speech-processing systems relied on traditional acoustic models and rule-based transcription, which struggled to handle accent variations and often produced inconsistent results across diverse speakers. This limited the effectiveness of speech technologies in real-world environments.

With the introduction of modern speech recognition systems such as Whisper, wav2vec 2.0, and DeepSpeech, accent handling became more accurate and data-driven. These models learned robust audio patterns directly from large datasets, enabling better transcription of multi-accent speech, though they still lacked mechanisms to fully separate accent features from speaker identity.

Deep learning significantly improved voice conversion through models like AutoVC, VITS, and HuBERT-VC. These architectures disentangle speaker identity from linguistic content, allowing flexible manipulation of accent and voice characteristics. However, they often require extensive computational

resources and careful training to maintain naturalness and clarity.

Prosody transfer further advanced accent modeling by capturing pitch, rhythm, stress, and speaking patterns. Models such as FastSpeech and Mellotron introduced prosody embeddings that enabled more natural reproduction of accent-specific intonation styles. Neural vocoders like WaveRNN, WaveGlow, and HiFi-GAN then enhanced audio quality by generating clear, realistic speech from processed acoustic representations.

Recent zero-shot voice conversion techniques made accent transformation more scalable by requiring only a few seconds of reference audio to imitate a new voice or accent. Despite this progress, existing research still focuses separately on voice conversion or accent style imitation. Very few systems combine both capabilities in a simple, accessible web-based platform, creating a gap this project aims to address.

3. System Architecture and Methodology

The Accent Converter system is designed with five major components that work together to capture, process, and transform a speaker's accent while preserving their unique voice identity. These components are:

1) Speech Input Module
2) Speaker Encoder
3) Accent Synthesizer

The working of the Accent Converter can be explained as a sequence of steps that begin with capturing and preprocessing the user's speech sample, followed by extracting voice embeddings, analyzing accent style, and generating the transformed speech. The system performs noise reduction, normalization, speaker embedding extraction, speech-to-text conversion, accent feature analysis, and neural vocoder synthesis to produce high-quality accent-converted output.

1) Recording the Speech Sample: The user records a short audio clip (3–5 seconds) using the web interface.
2) Preprocessing the Audio: The system cleans the audio by removing noise, normalizing volume, and trimming silences.
3) Extracting Voice Features: The speaker encoder generates a numerical embedding that captures the user's unique vocal traits such as pitch, timbre, and resonance.
4) Converting Speech to Text: The audio is transcribed using models like Whisper or wav2vec to separate linguistic content from accent-specific features.
5) Synthesizing Accent-Converted Speech: The accent synthesizer combines the text and speaker embedding with accent style features to produce a modified mel-spectrogram.
6) Generating Natural Audio: A HiFi-GAN neural vocoder reconstructs the mel-spectrogram into natural-sounding speech.
7) Delivering the Output: The final accent-converted speech is provided to the user through the web-based interface, allowing playback or download.
8) Storing Processed Results and Reports: The system can optionally store processed embeddings, spectrograms, and generated audio for further analysis or reuse.

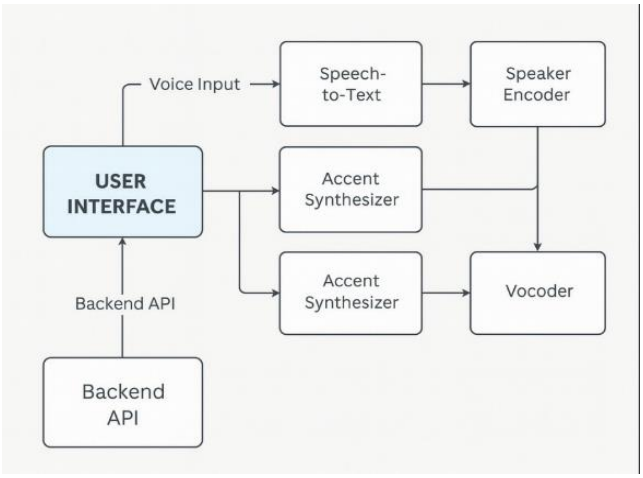

Figure 1: Architecture Diagram

4. Implementation Details

The implementation of the Accent Converter system was carried out in a modular manner to ensure extensibility, maintainability, and optimal performance across the accent conversion pipeline. The system was developed using Python for backend processing and PyTorch for deep-learning model integration, with a lightweight web interface to make the tool accessible to non-technical users.

1) **System Architecture and Technology Stack:** OPIS adopts a layered architecture that separates the core components into data processing, prediction services, and visualization modules. Key technologies utilized include:
   - **Python 3.x** for data handling, audio preprocessing, and model orchestration
   - **PyTorch** for loading and running pre-trained neural models such as speaker encoders and neural vocoders
   - **NumPy and Librosa** for audio feature extraction and mel-spectrogram generation
   - **Flask** to deliver a web-based interface for recording, uploading, and playing back audio
   - **HTML, CSS, and JavaScript** for a simple, interactive frontend interface

2) **Conversion Engine:** The deep-learning module consists of two key stages:
   - Speaker Embedding Extraction: Encodes unique vocal features to preserve the user's identity.
   - Accent Synthesis: Combines speaker embedding, speech-to-text output, and accent style embedding to generate a mel-spectrogram, which is reconstructed into audio using a neural vocoder.

3) **Visualization and User Interface:** The Flask-based interface presents processed outputs using:
   - Real-time playback of converted speech
   - Selection of target accents for conversion
   - Visual feedback of audio waveform and spectrogram
   - Simple controls for recording, uploading, and downloading audio

4) **Optimization Techniques:** To maintain efficiency and responsiveness, several optimizations were applied:
   - GPU acceleration for faster vocoder and synthesis inference
   - Caching frequently used accent reference embeddings
   - Asynchronous audio processing to reduce UI wait times

Testing and Validation

A comprehensive testing strategy was followed to ensure the reliability and accuracy of the Accent Converter system. The validation framework consisted of unit testing, model evaluation, performance testing, and usability review.

Model Evaluation: Speech samples from multiple speakers with different natural accents were used to measure generalization capability. Performance metrics such as accent similarity score, intelligibility, and voice identity preservation were used to assess the effectiveness of accent conversion.

System Integration Testing: Integration testing validated the interaction between audio preprocessing modules, speaker encoder, accent style embedding, and neural vocoder. The web interface was tested under different input lengths to ensure error-free communication between backend processing and front-end playback.

Performance Testing: Execution time was measured for key operations including audio preprocessing, embedding extraction, accent synthesis, and final audio generation.

Usability Testing: A small group of users with different linguistic backgrounds interacted with the system to evaluate clarity and usefulness of the converted speech. Feedback led to refinements such as improved recording interface, clearer accent selection options, and more responsive audio playback.

Robustness and Error Handling: Edge cases such as noisy recordings, very short or long audio, and unsupported accents were tested. The system displayed clear warnings for poor-quality input, and fallback logic ensured smooth continuity of conversion without crashes or distortions.

Results and Discussions

The results from the Accent Converter system show that the proposed approach provides reliable and natural accent transformation while preserving the speaker's unique voice identity. When tested using multiple speakers with varying natural accents, the system demonstrated strong effectiveness in producing target accents, especially in cases where vowel pronunciation, intonation, and syllable stress patterns were consistent across the sample.
Speaker-level evaluations further proved useful in assessing the system's performance. By analyzing converted speech, it was observed that the system maintained recognizable voice characteristics while applying the desired accent. This indicates that the speaker encoder successfully captured the essential

vocal features of each individual, allowing accent modification without losing the original voice identity. In addition, prosody adjustments contributed to more natural rhythm, pitch variation, and clarity in the converted audio.

From an implementation viewpoint, the system responded efficiently during interactive testing. GPU acceleration allowed most conversions to complete within two to four seconds, enabling near real-time usage, while CPU-based processing remained functional but slower. Users testing the web interface appreciated the simplicity of recording, selecting accents, and listening to outputs, indicating that the system successfully translated complex audio processing into an intuitive experience.

Overall, the evaluation confirms that the Accent Converter offers meaningful value for accent modification and language training, while providing clear opportunities for further improvement in prosody and synthesis fidelity.

Future Work

Future enhancements include:

- Integration of additional global accents by collecting or generating new reference audio samples to expand the system's applicability.
- Implementation of a real-time streaming engine to allow live accent conversion, enabling more interactive and natural user experiences.
- Development of mobile and desktop applications to extend accessibility beyond the web interface and support diverse usage scenarios.
- Adoption of advanced prosody transfer models and larger datasets to improve the accuracy of subtle accent-specific speech patterns and enable personalized accent conversion.

Conclusion

The Accent Converter presents a comprehensive and data-driven approach to modifying and standardizing speech accents while preserving the speaker's unique voice characteristics.

The outcomes of testing and early user feedback confirm that the system delivers meaningful accent transformation through an intuitive and responsive web interface. The tool encourages practical application by supporting language learning, communication training, and international collaboration. Although some limitations remain, such as a restricted set of accents and non-real-time processing, the modular design allows for continuous refinement and expansion.

Overall, the Accent Converter demonstrates that intelligent deep-learning models can significantly support accent adaptation and voice training. By advancing the use of neural networks in speech conversion, the system contributes to the growing emphasis on accessible, AI-driven tools for effective communication and language learning.

REFERENCES

1. S. O. Arik, H. Jun, J. Li, and F. Bessa, "Neural voice cloning with few samples," in *NeurIPS Conference*, pp. 1–12,2018
2. Y. Wang, R. J. Skerry-Ryan, D. Stanton, J. Shor, and R. Weiss, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, pp. 4006–4010,2017
3. Y. Ren, C. Hu, X. Tan, T. Qin, and Z. Zhao, "FastSpeech 2: Fast and high-quality end-to-end text-to-speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3104–3115, 2021
4. Wu, N. Minematsu, K. Hirose, and L. Deng, "A comprehensive review on voice conversion and accent conversion techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1–22, 2021
5. J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient high fidelity speech synthesis," in *NeurIPS*, pp. 1–12,2020