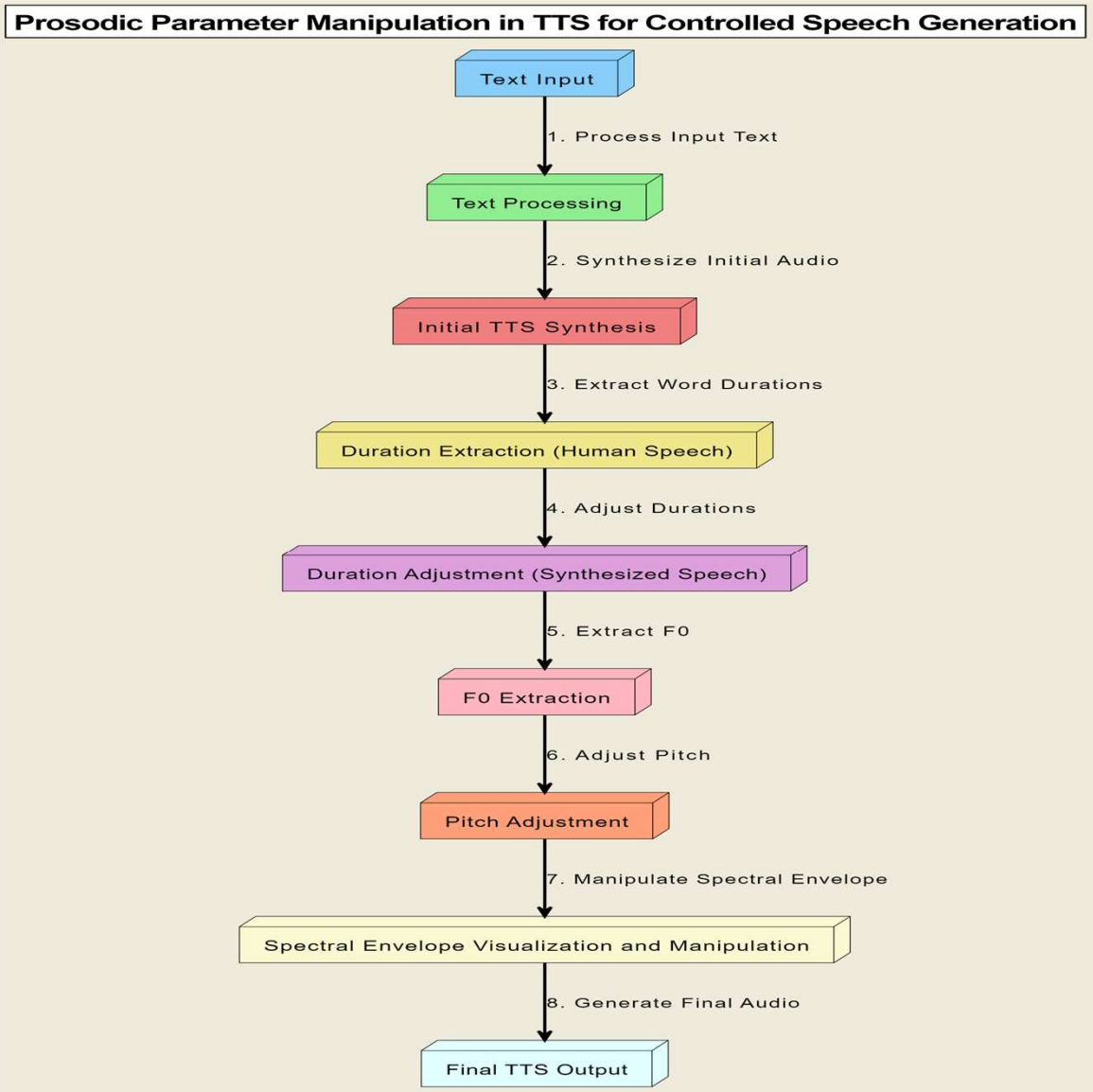
Prosodic Parameters manipulation in TTS for controlled speech generation

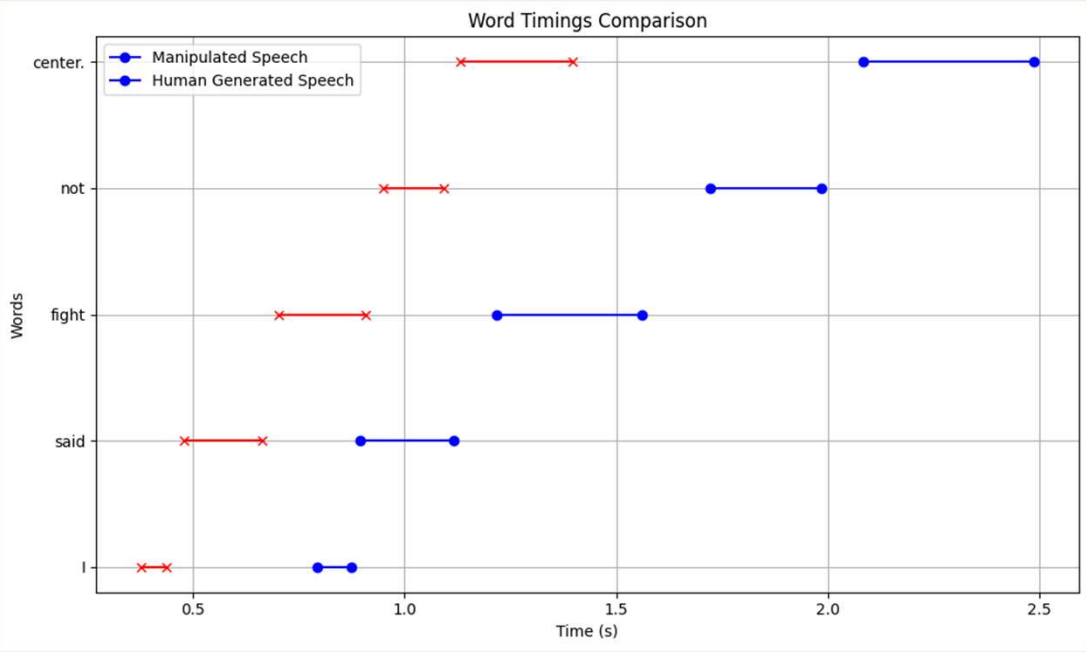
Trizen

Aim to enhance speech expressiveness and emotional quality through prosodic manipulation

Fundamental Frequency (F0)
Spectral Envelope
Duration
Intensity (Loudness)
Tempo (Speech Rate)


Prosodic Parameter Manipulation in TTS for Controlled Speech Generation
Text Input
1. Process Input Text
Text Processing
2. Synthesize Initial Audio
Initial TTS Synthesis
3. Extract Word Durations
Duration Extraction (Human Speech)
4. Adjust Durations
Duration Adjustment (Synthesized Speech)
5. Extract F0
F0 Extraction
6. Adjust Pitch
Pitch Adjustment
7. Manipulate Spectral Envelope
Spectral Envelope Visualization and Manipulation
8. Generate Final Audio
Final TTS Output

Results

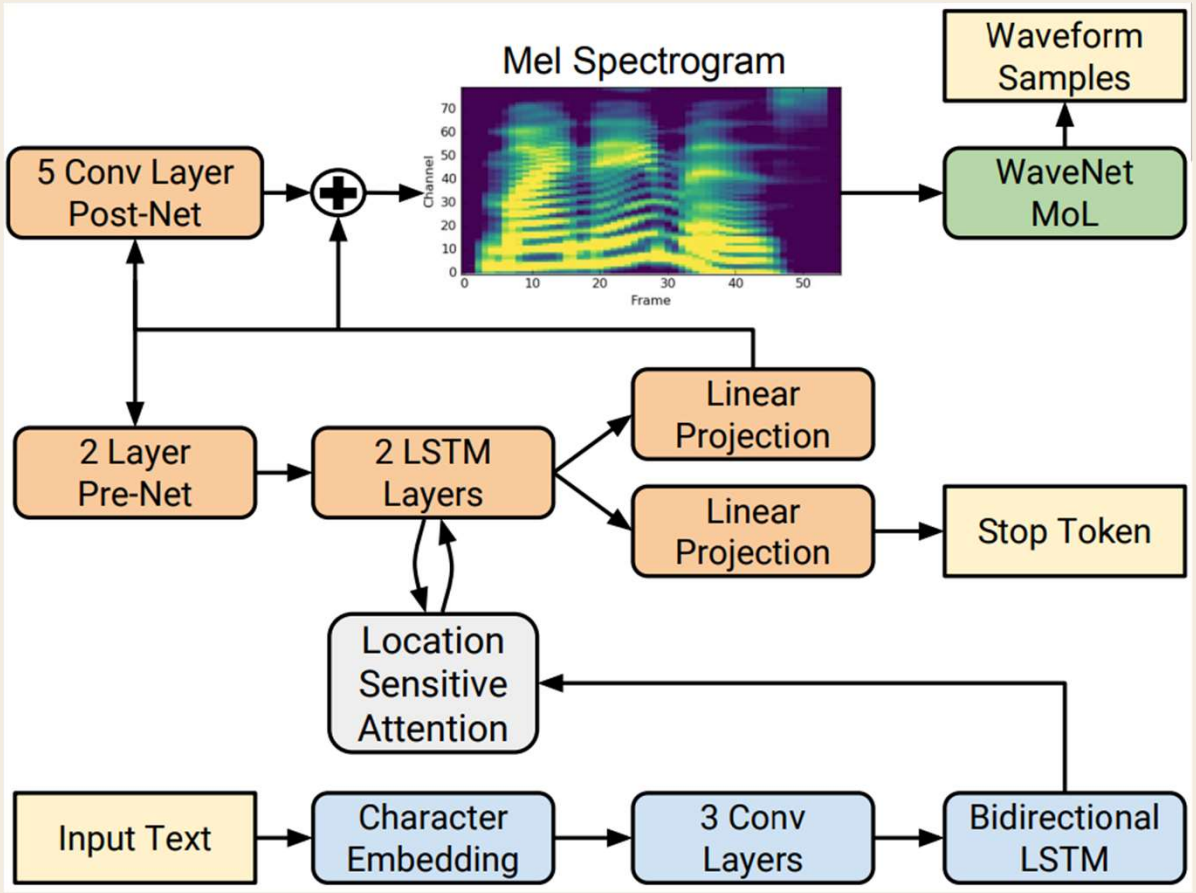

Word Timings Comparison
Manipulated Speech
Human Generated Speech
center.
not
fight
said
I
Words
0.5
1.0
1.5
2.0
2.5
Time (s)

Prosodic Parameter Manipulation in TTS Generated Speech for Controlled Speech Generation

AT SENTENCE LEVEL

TEXT – TO –SPEECH (TTS)

TACOTRON 2



Extracting Essential Features

Fundamental Frequency (F0)

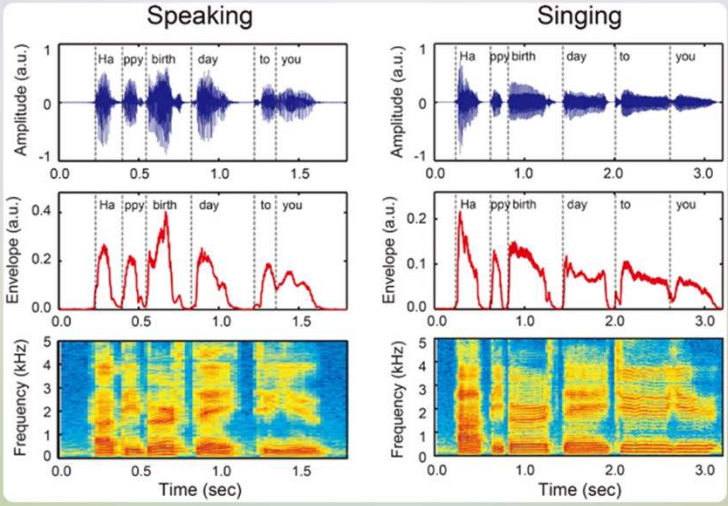
The fundamental frequency, or F0, represents the vibration rate of the vocal cords, which is a crucial aspect of perceived pitch. Accurately extracting and modifying the F0 is essential for adjusting the prosody and intonation of synthesized speech.

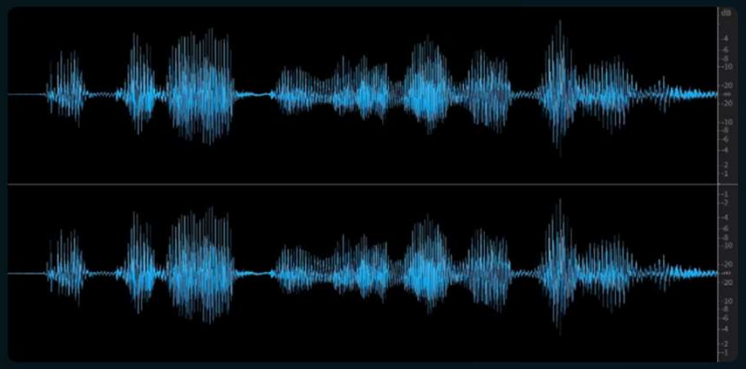
Spectral Envelope (SP)

The spectral envelope, or SP, captures the overall shape of the audio spectrum, providing information about the timbre and quality of the voice. Scaling the SP allows for adjustments to the energy and resonance characteristics of the synthesized speech.

Aperiodicity (AP)

Aperiodicity, or AP, measures the degree of randomness or non-periodic components in the audio signal. This feature is critical for preserving the natural, expressive qualities of the voice, especially during unvoiced segments like consonants.





Modifying and Synthesizing Speech

1

Pitch Adjustment

The pitch of the synthesized speech is modified by scaling the fundamental frequency (F0) while preserving the natural contour of the pitch trajectory. This allows for adjustments to the perceived intonation and expressiveness of the voice.

2

Duration Adjustment

The duration of the speech features (F0, SP, AP) is modified using time-stretching techniques. This enables control over the pacing and rhythm of the synthesized utterance, helping to align it more closely with natural human speech patterns.

3

Energy Adjustment

The energy, or volume, of the synthesized speech is adjusted by scaling the spectral envelope (SP). This helps to balance the perceived loudness and emphasis of the synthetic voice, ensuring it matches the desired expression and emphasis.

METHODOLOGY

- Pitch Manipulation: The modify_pitch_preserve_contour function adjusts the pitch while preserving its overall contour: It works in the log domain to maintain relative pitch differences. It normalizes the pitch contour, applies the shift, and then denormalizes it. This preserves the natural rises and falls in pitch while adjusting the overall pitch level.

- Energy Manipulation: The modify_energy function simply multiplies the spectral envelope by an energy factor: ***return sp * energy_factor*** This scales the energy of the speech uniformly across all frequencies.

- Duration Manipulation: This uses interpolation to resize the feature arrays and also use Duration Factor, whether to Speed Up or Slow Down it.

- Combined Manipulation: The manipulate_features function applies all these manipulations together: It modifies the pitch (f0) while preserving the contour. It adjusts the duration of f0, spectral envelope (sp), and aperiodicity (ap).It modifies the energy of the spectral envelope.

Automating the Workflow

1

File Preparation

The system starts by setting up the necessary paths to the human-recorded and TTS-generated audio files, as well as creating output directories to store the processed results.

2

File Processing

The system then iterates through each pair of human-recorded and TTS-generated audio files, processing them using the detailed feature extraction, modification, and synthesis steps.

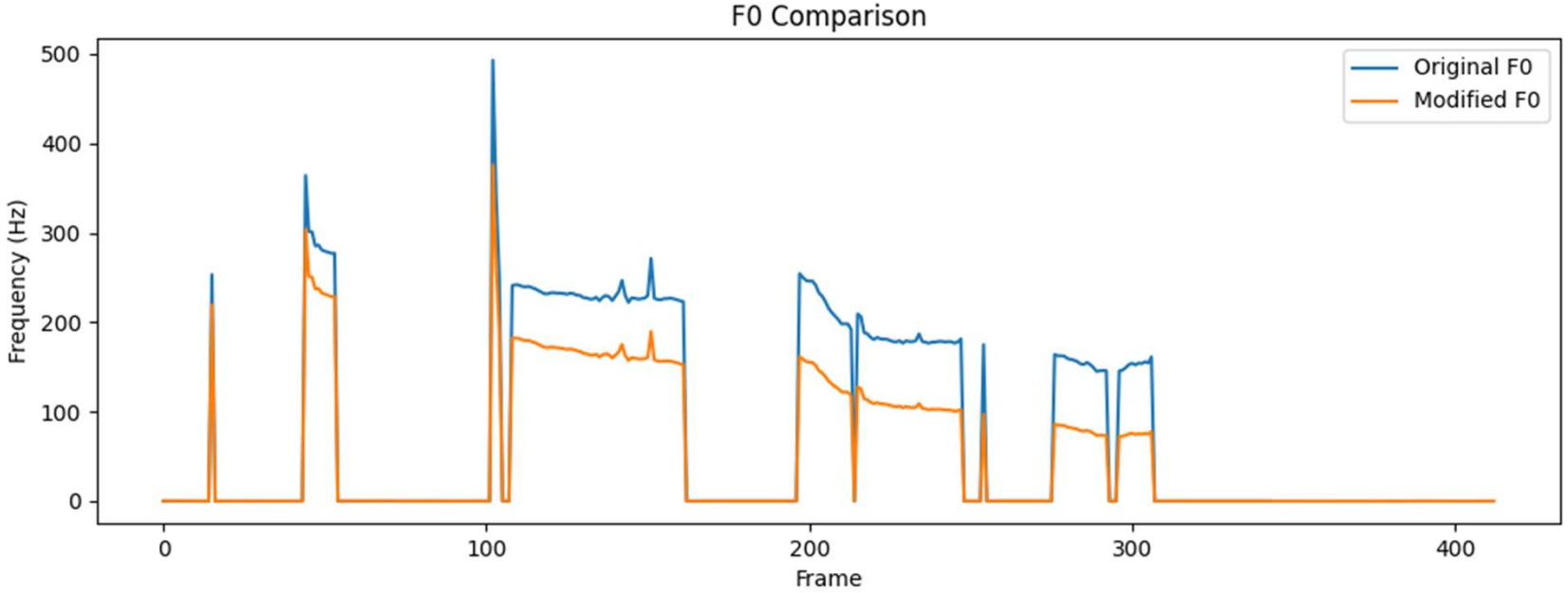
3

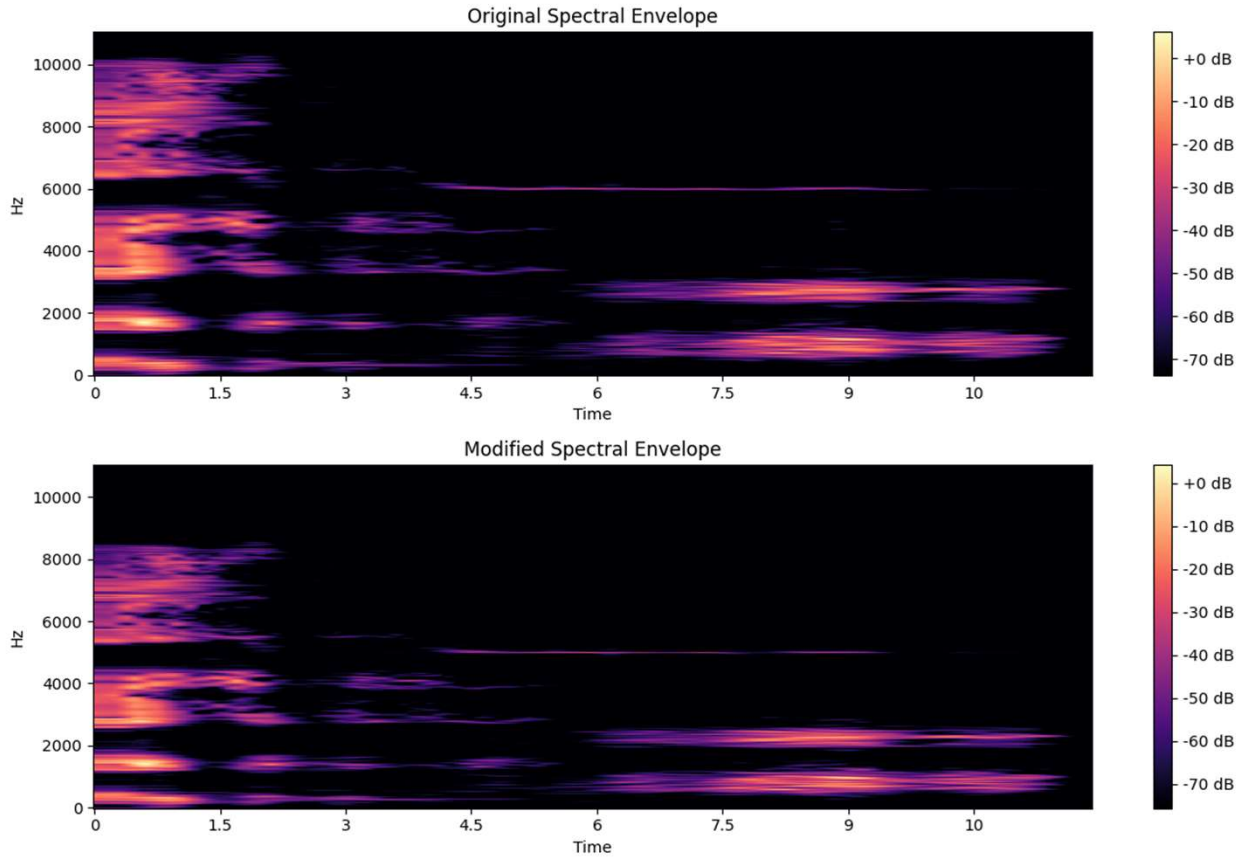
Comparison and Optimization

Finally, the system compares the features of the human-recorded and modified TTS-generated audio, providing insights to further refine and optimize the synthetic voice for improved expressiveness.

BASIC OUTPUT

TEXT – I SAID SNOW NOT TOMORROW

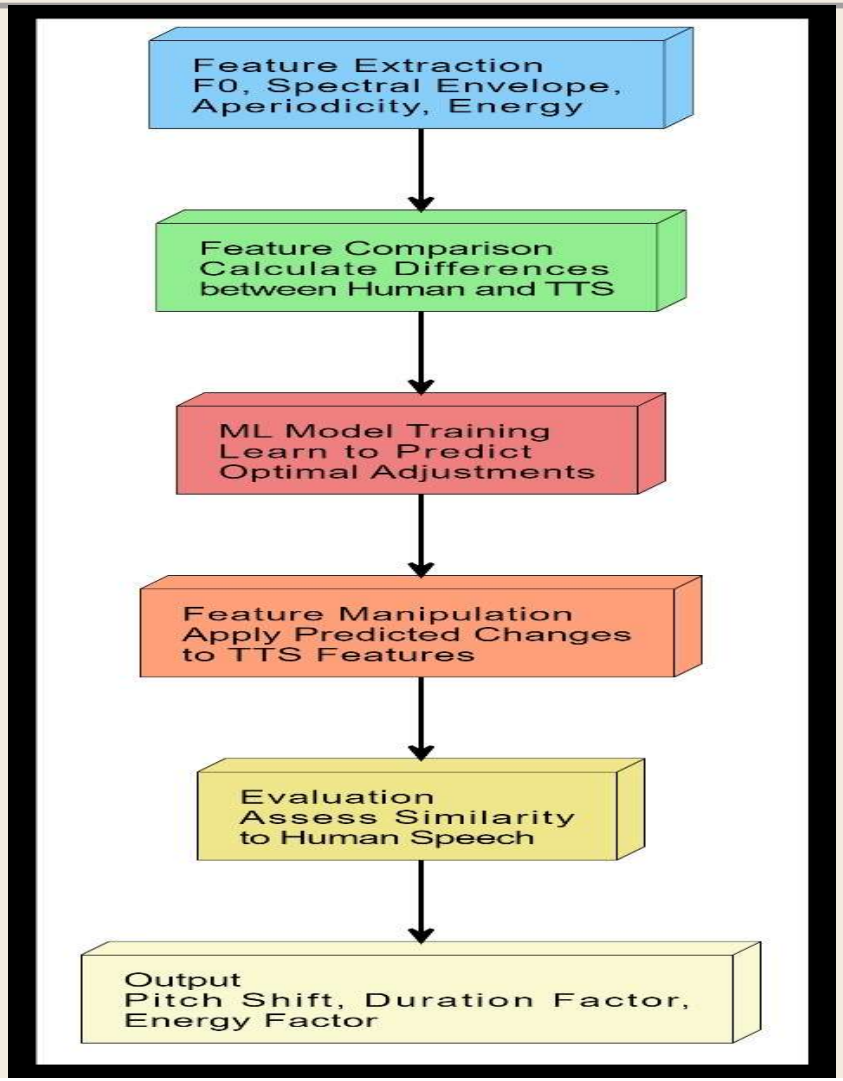

F0 Comparison
Original F0
Modified F0
Frequency (Hz)
Frame
0
100
200
300
400
500


Original Spectral Envelope
Hz
10000
8000
6000
4000
2000
0
0
1.5
3
4.5
6
7.5
9
10
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
Modified Spectral Envelope
Hz
10000
8000
6000
4000
2000
0
0
1.5
3
4.5
6
7.5
9
10
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB

Overview

- Goal: Make TTS speech sound more natural and human-like
- Method: Adjust pitch, duration, and energy using ML techniques
- Process: Extract prosodic features from human and TTS speech
- Compare features to identify differences
- Train ML model to predict optimal adjustments
- Apply predicted changes to TTS audio
- Evaluate similarity to human speech

Flowchart


Feature Extraction
F0, Spectral Envelope,
Aperiodicity, Energy
Feature Comparison
Calculate Differences
between Human and TTS
ML Model Training
Learn to Predict
Optimal Adjustments
Feature Manipulation
Apply Predicted Changes
to TTS Features
Evaluation
Assess Similarity
to Human Speech
Output
Pitch Shift, Duration Factor,
Energy Factor