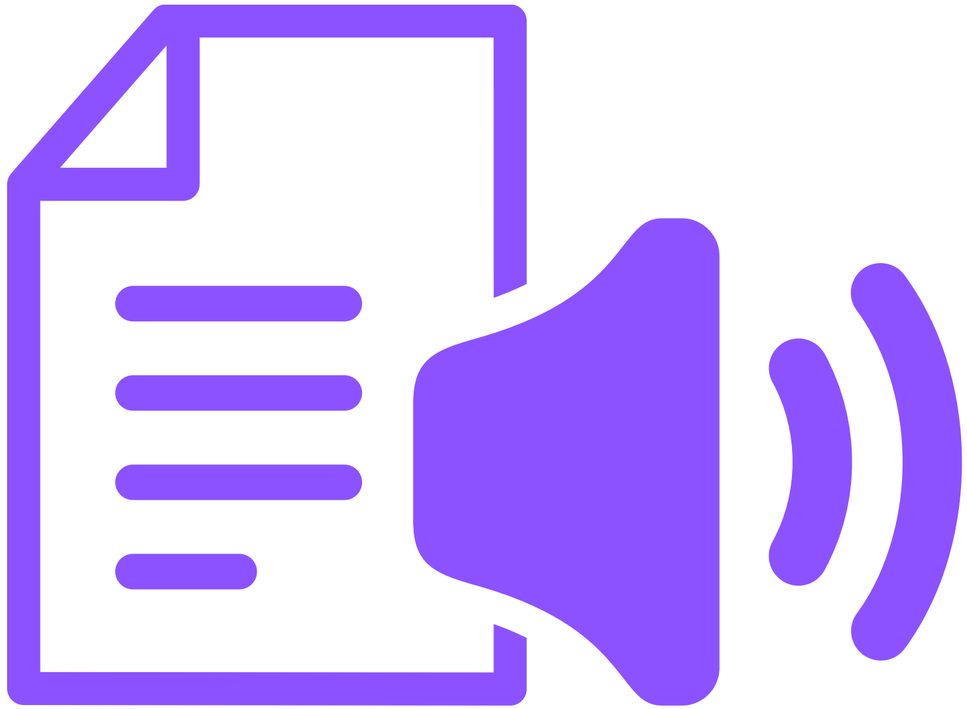
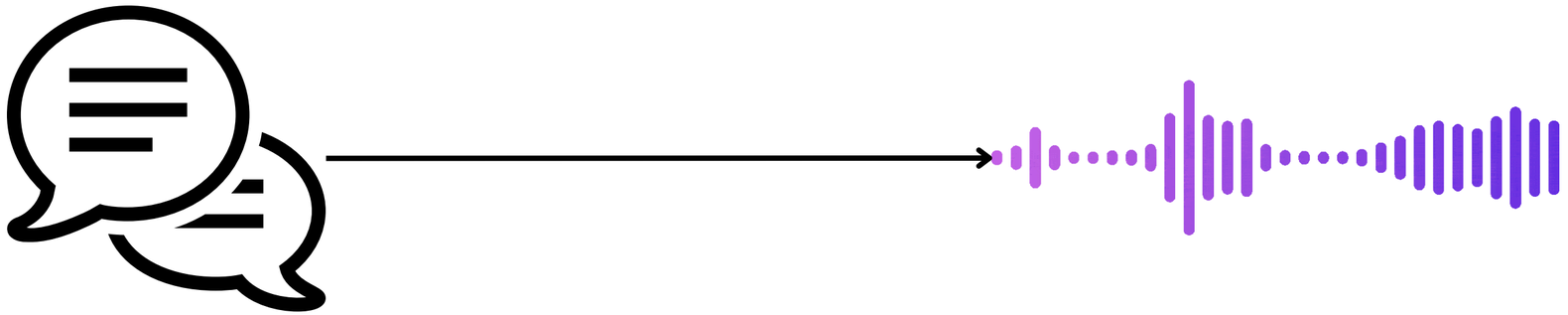


Text to Speech Synthesis

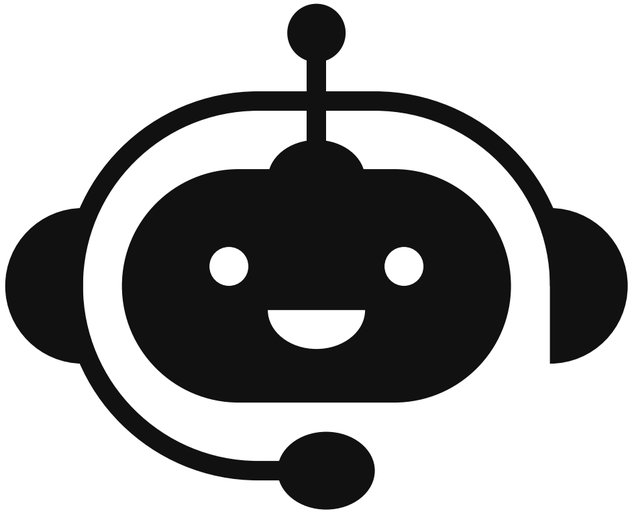
Presented by: Yassine ibork
Supervised by: Dr. Kshirsagar Shruti

What is text to speech?

Text-to-Speech (TTS) is a technology that converts written text into spoken words. It allows computers and devices to generate human-like speech, enabling users to listen to written content instead of reading it.

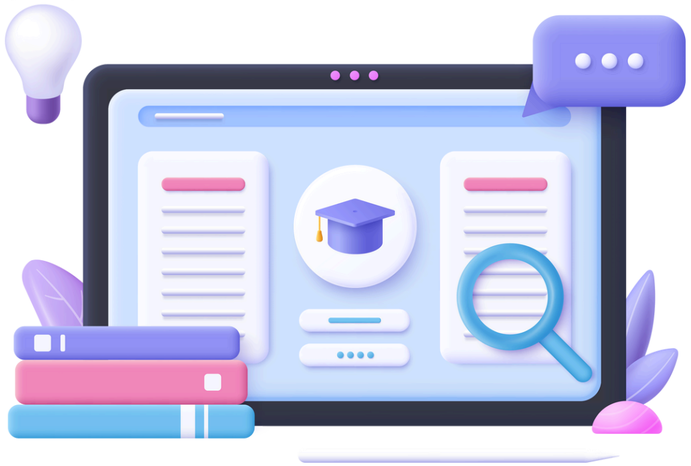


Use Cases of Text-to-Speech



Voice Assistants

Navigation Systems



E-Learning

TODAY'S AGENDA

- TERMINOLOGIES OF TTS SYSTEMS
- HOW DOES TTS WORK?
- EVOLUTION OF TEXT TO SPEECH SYSTEMS
- EXPLORING TTS SOLUTIONS: ON-PREMISES VS. CLOUD
- HANDS-ON DEMO
- CONCLUSION

I. Terminologies of TTS systems

Phoneme

The smallest unit of sound in a language that distinguishes one word from another (e.g., the "b" in "bat").

Prosody

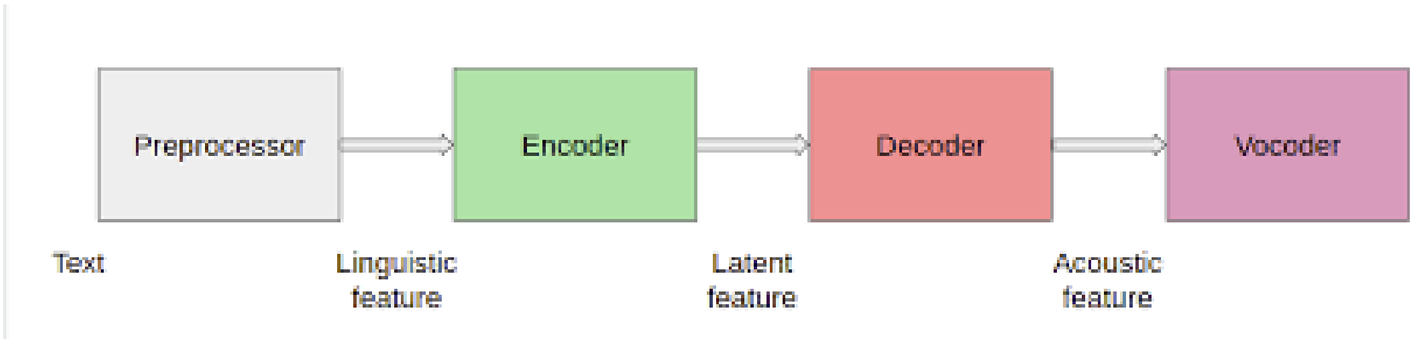
Refers to the rhythm, stress, and intonation of speech, which helps convey emotions and natural flow.

Mel-Spectrogram

A visual representation of sound, showing the distribution of energy across different frequencies over time, used by TTS systems to convert text into speech.

II. How does TTS work?

Text-to-Speech (TTS) systems convert written text into human-like speech using a multi-step process. This involves:



Preprocessor

Tokenization

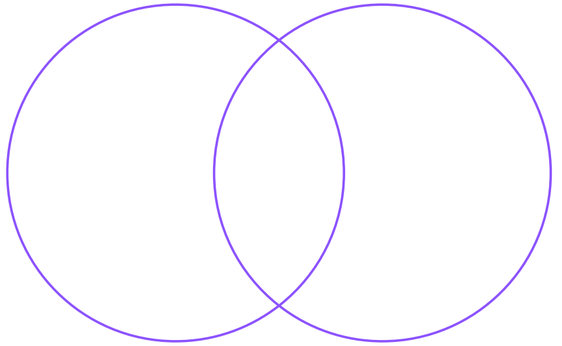
Breaks the input text into smaller units, like words or sentences.

Phoneme Conversion

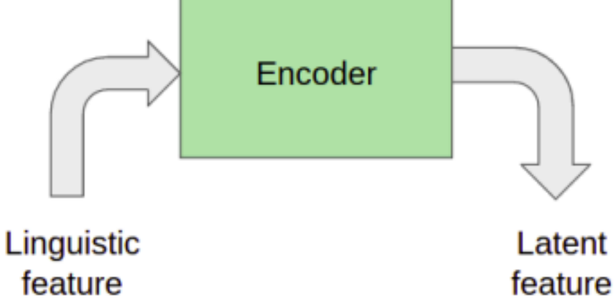
Converts text into phonemes based on pronunciation (e.g., "hello" becomes HH AH0 L OW1).

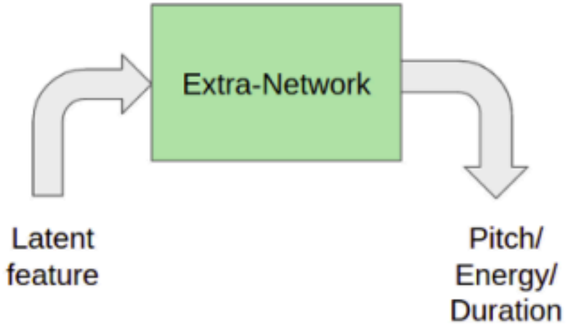
Phoneme Duration

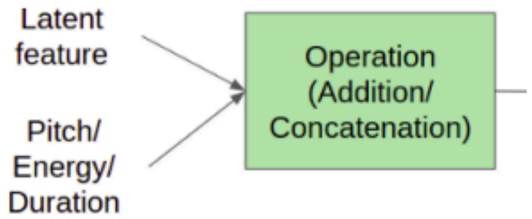
Calculates the time each phoneme takes in the audio, affecting the speech's pacing.



Encoder







Phoneme Input

The encoder receives phonemes (linguistic features) as input.

Feature Extraction

Transforms phonemes into n-dimensional embeddings, capturing essential speech features.

Latent Features

These embeddings, known as latent features, are crucial for predicting prosody (pitch, energy, duration) and improving the naturalness of speech.

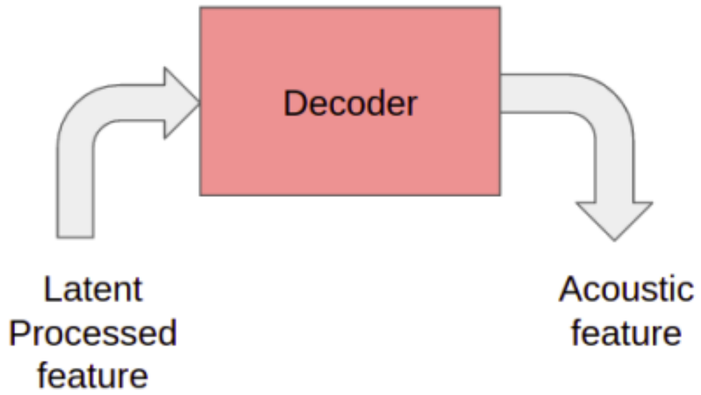
Decoder

Latent Feature Input

The decoder takes the latent features generated by the encoder.

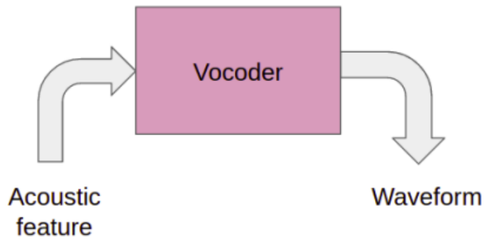
Mel-Spectrogram Generation

It converts these features into a mel-spectrogram, representing the frequency and amplitude of the speech signal over time.



Vocoder

It transforms the acoustic features (Mel-spectrogram) into a waveform output (audio). This can be achieved using a mathematical model like **Griffin Lim**, or by training a neural network to map Mel-spectrograms to waveforms. In practice, learning-based methods generally outperform the Griffin Lim model.

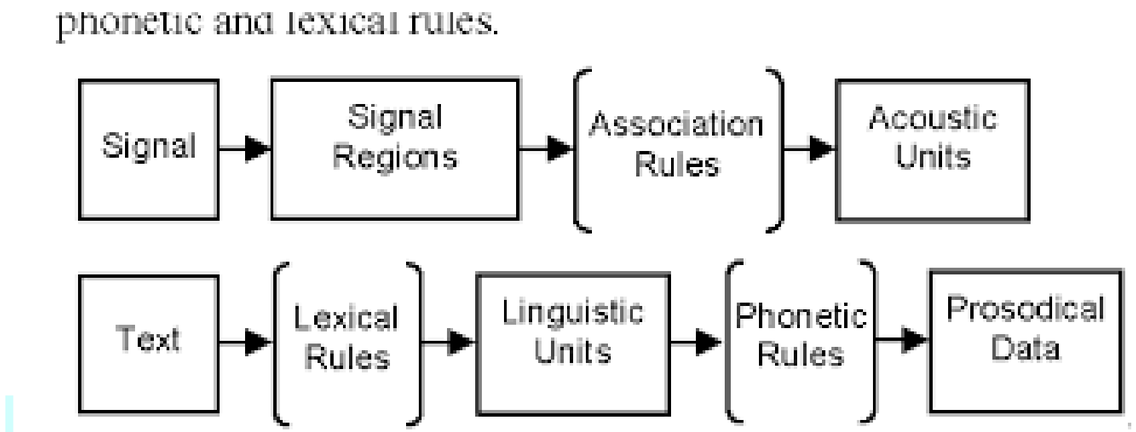


III. Evolution of text to speech systems

1. Rule-based Systems
2. Concatenative TTS
3. Statistical Parametric TTS
4. Neural TTS

Rule-based Systems

- **A rule-based Text-to-Speech (TTS) system is a traditional method for converting written text into spoken language using predefined linguistic and phonetic rules.**
  - **Lexical Rules: Identify the structure of the text.**
  - **Linguistic Units: Convert the text into fundamental speech units.**
  - **Phonetic Rules: Apply the pronunciation rules to convert linguistic units into phonetic forms.**
  - **Prosodic Data: Add intonation, rhythm, and stress to generate natural-sounding speech.**

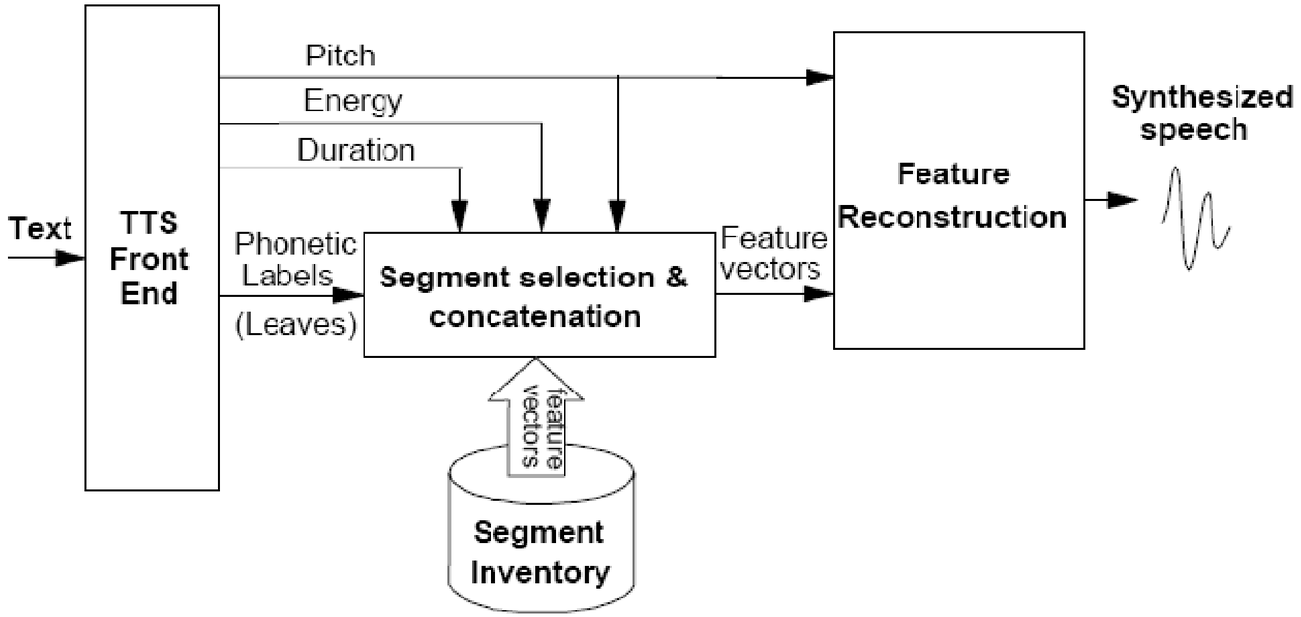


Limitations of rule based tts

- **Robotic or Monotone Output:** Rule-based TTS systems produce sound that lacks emotional expressiveness, making them sound robotic.
- **High Manual Effort:** Rule-based systems require linguists and engineers to manually craft rules for pronunciation, grammar, prosody, and sentence structure.
- **Scalability Challenges:** Scaling rule-based systems to support multiple languages, accents, or styles requires significant manual intervention, leading to scalability issues.

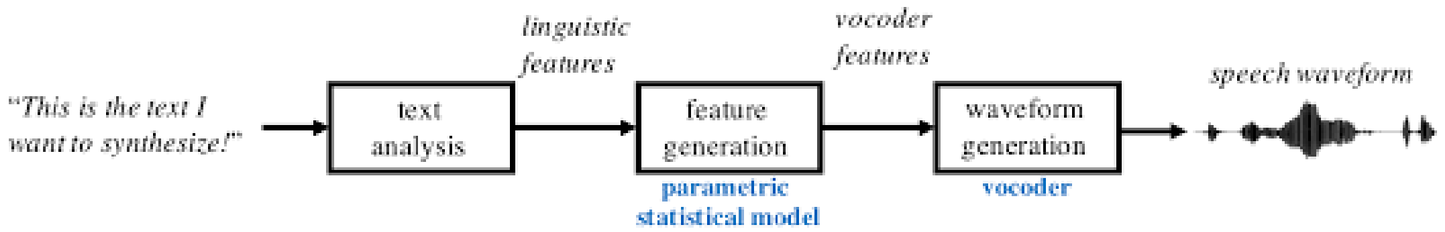
Concatenative TTS

- Concatenative Text-to-Speech (TTS) is a method of speech synthesis that constructs speech by selecting and concatenating pre-recorded segments of human speech (e.g., phonemes, diphones, syllables).
- By using actual recordings of human speech, concatenative TTS systems produce much more natural-sounding output compared to rule-based systems, which rely on generated waveforms or acoustic models.



Statistical Parametric TTS

- Parameter Modeling: The core of this approach is modeling speech parameters (pitch, spectral features, duration) using statistical techniques. This allows for more flexible speech generation compared to concatenative methods.
- Training from Dataset: The system learns from a training dataset of recorded speech, typically from a single speaker. This enables it to capture the characteristics of natural speech.
- Statistical Models: It uses statistical models (like **Hidden Markov** Models in early versions) to represent sounds or phonemes as probability distributions over speech features. This probabilistic approach allows for smoother transitions between sounds.
- Flexibility: Unlike concatenative TTS, this approach can potentially generate speech for any text input, even if specific word combinations weren't in the training data.



Neural TTS

- Utilizes deep learning models to directly generate waveforms from text.
- Significantly improves naturalness, expressiveness, and flexibility, resulting in human-like voices.
- Examples: Tacotron, WaveNet.

IV. Exploring TTS Solutions: On-Premises vs. Cloud

- 1. **On-Premises TTS:** Leveraging Pre-trained Models
  - **Control and Privacy:** Implement TTS locally, ensuring full control over data.
  - **Pre-trained Models**: Use libraries like SpeechBrain for quick setup and high-quality speech synthesis without extensive ML knowledge.
- 2. **TTS as a Service:** Cloud Providers
  - **Google Cloud Text-to-Speech:**
    - Multiple voice options and language support.
    - Easy API integration for applications.
  - **Amazon Polly:**
    - Natural-sounding voices with real-time synthesis.
    - Integrates seamlessly with other AWS services.

Hands-On Demo

Code Available at: https://github.com/yibork/spoken-language-processing-tts-presentation

IV. Conclusion

Text-to-Speech systems have evolved from rule-based approaches to advanced neural architectures, providing more natural and expressive speech. Neural TTS, with its deep learning capabilities, has revolutionized speech synthesis, enabling customizable voices, real-time generation, and significant improvements in quality and flexibility for various applications.

References

- A review-based study on different Text-to-Speech technologies
- Text To Speech for Bangla Language using Festival