Applying Hidden Markov Models (HMM) in African-Accented Text-to-Speech (TTS) Systems

Capstone Context

My capstone project focuses on improving the accessibility of digital content through a fine-tuned text-to-speech (TTS) system tailored to East and West African users. The goal is to generate natural, intelligible speech that accurately reflects local phonetic and prosodic patterns. Hidden Markov Models (HMMs), though largely replaced by deep learning in modern systems, still offer a structured way to model speech synthesis—especially in resource-constrained or interpretable environments.

Observations

In this TTS context, observations are the **acoustic features** extracted from speech recordings, such as:

- Mel-Frequency Cepstral Coefficients (MFCCs)
- Pitch contours (F0)
- Duration of phonemes
- Energy levels

These features are generated from aligned training data and used as output signals to be predicted given a sequence of text units (e.g., phonemes or graphemes).

Type of HMM Problem

This is a **supervised sequence modeling task**, where the **hidden states (e.g., phonemes)** are known during training. The task is to learn how to generate the observed acoustic signals from these states. The system is trained to model the **alignment and generation** process between phoneme sequences and their corresponding audio features.

Training Algorithm

a. Known Values at the Start:

- Labeled training data: pairs of phoneme sequences and aligned acoustic features
- Number of states (typically one per phoneme, often expanded into sub-states)

b. Unknown Values to Learn:

- Transition probabilities between states (e.g., phoneme durations)
- Emission probabilities from each state (mapping to audio features)

The **Baum-Welch algorithm** is used to iteratively estimate these unknown parameters from data.

Parameter Updates

The HMM training process updates:

- **Transition matrix**: Probabilities of progressing through phoneme sub-states or transitioning to new phonemes
- **Emission matrix**: Distribution of acoustic features given a state
- **Initial state probabilities**: Likelihood of starting in a particular phoneme or sub-state

This structured modeling can support baseline synthesis and be used in hybrid systems where HMMs handle low-resource edge cases, dialect adaptation, or fallback pronunciation models. It enhances intelligibility and inclusiveness for African users, even in data-sparse environments.