Secure Linux Voice Command Controller

Libin N George

Indian Institute of Technology Palakkad

111501015@smail.iitpkd.ac.in

October 16, 2018

Overview

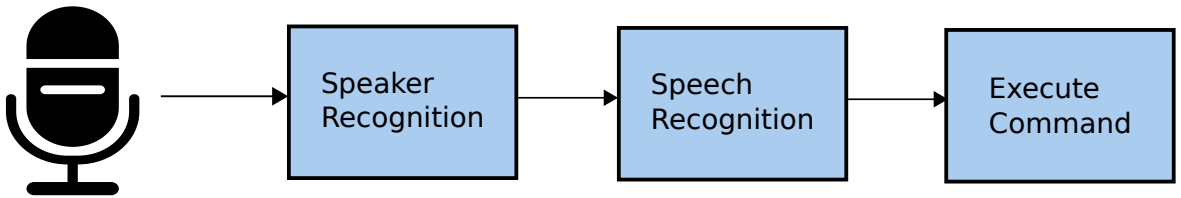
1. Intoduction
   - Introduction
   - Project Goal
   - RoadMap
2. Classes of Speech Signal
   - Voiced Speech
   - Unvoiced Speech
   - Non Speech(Silence)
3. Voice Activity Detection
4. Speech Recognition
   - Hidden Markov Model
5. References

Intoduction

Operating systems like Windows and Mac OS contains voice controller which allows users to have hands-free control of Computers. In this project, we are creating a voice controller for Linux systems which can recognize and execute voice commands like the commercial virtual assistants. For example, a voice command open firefox should open firefox.

Project Goal

Developing a secure voice-based controller for Linux desktops. The controller should recognize the speaker and respond only if the speaker is registered in the system.



Roadmap

Phase 1

- **Step 1.1** : Voice Activity Detection(VAD).
  Voice Activity Detection(VAD) helps in reducing the load in the system by stopping audio processing when not required.
- **Step 1.2** : Speech Recognition.
  Speech recognition should be capable of recognising at least some words (for which we trained the system).

Phase 2

- **Step 2.1** : Speaker Recognition. We need the voice controller to be secure. That is, the application should be able to recognize the speaker (owner) and should not execute commands from other people.
- **Step 2.2** : Executing commands on speech recognition.
- **Step 2.3** : Creating UI and Testing.

Classes of Speech Signal [2]

- Voiced Speech
- Unvoiced Speech
- Non-Speech (Silence)

Voiced Speech

Voiced Speech looks like a periodic signal. During the production of voiced speech, the air breathes out of lungs through the trachea is interrupted periodically by the vibrating vocal cords. Due to this interruption, glottal waves are generated which excites the speech production a system to produce voiced speech. Voiced speech can be identified by the periodic nature of the speech waveform, the relatively high energy associated with the speech waveform.

Classes of Speech Signal

Unvoiced Speech

Speech which looks like random noise without any periodic nature and is termed as unvoiced Speech. During the production of unvoiced speech, the air breathes out of lungs through the trachea is not interrupted by the vibrating vocal cords. However, obstruction of air flow occurs scarcely from glottis somewhere along the length of the vocal tract to produce total or partial closure. Due to this modification of air flow, a stop or friction excitation occurs creating unvoiced speech.

Non-Speech (Silence)

Non-speech (silence) is present between voiced and unvoiced speech.

Voice Activity Detection Features [1]

- **Short-Term Energy**
  The energy associated with voiced speech is large when compared to unvoiced speech. Silence speech will have least or negligible energy when compared to unvoiced speech. Hence, Short Term Energy can be used for voiced, unvoiced and silence classification of speech.
- **Most Dominant Frequency Component**
  The voiced speech of a typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz.
- **Spectral Flatness Measure**
  Spectral flatness provides a way to quantify how noise-like a sound is, as opposed to being tone-like. The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum. The ratio produced by this calculation is often converted to a decibel scale for reporting, with a maximum of 0 dB and a minimum of $-\infty$ dB.

Voice Activity Detection Features

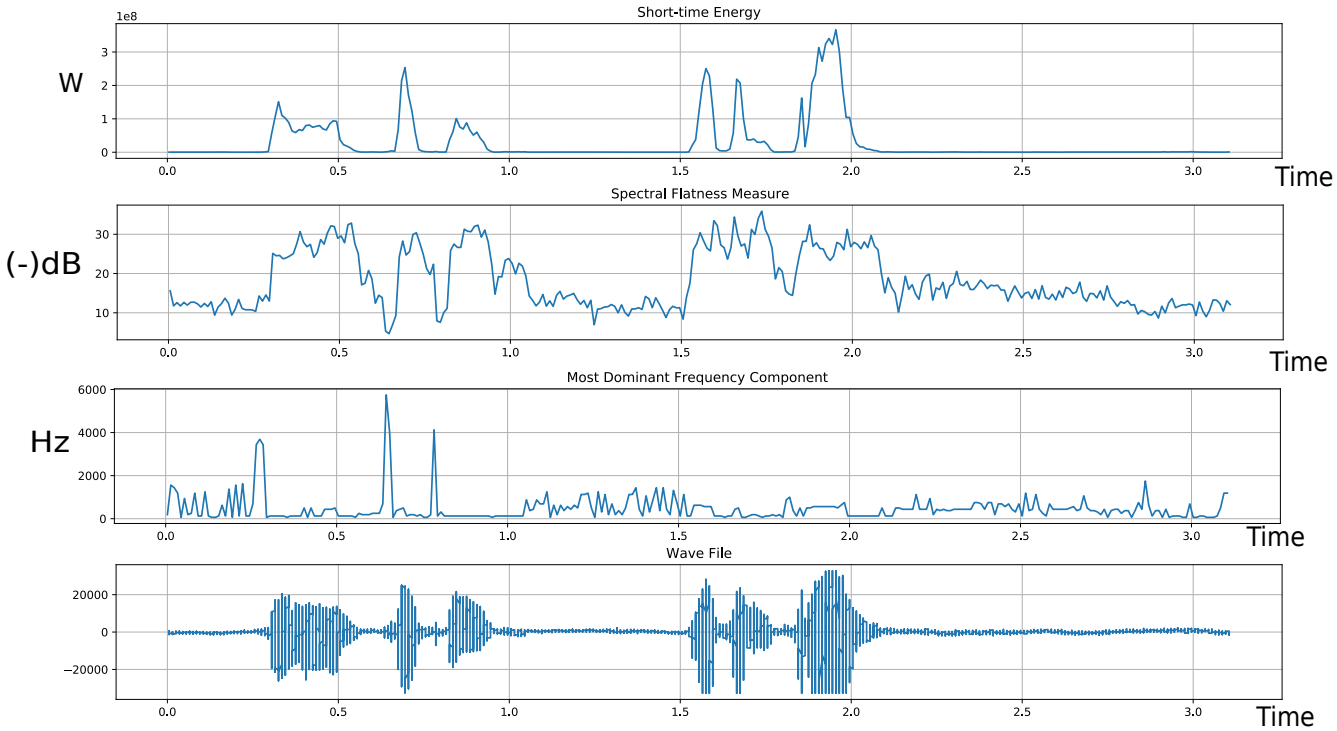


Figure: Experiment Results for different Short Term Features

Speech Recognition

Methods for Speech Recognition

- Using RNN and spectrogram
  We have to create a Spectrogram for the sound. Now recurrent neural network (RNN) can be used to recognize patterns in the Spectrogram.
- Using Hidden Markov Model.
  HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scale (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes. They also can be trained automatically and are simple and computationally feasible to use.

Hidden Markov Model

HMM is characterized by following

1. Number of State N
2. Number of Distinct Observation symbol per state M. The Observations are $\{V_1, V_2, \ldots, V_M\}$
3. State Transition Probability $a_{ij} = P[q_{t+1} = S_i | q_t = S_j], 1 \leq i, j \leq N$
4. Observation symbol probability distribution in state j, $B_j(K) = P(V_k \ at \ t | q_t = S_j)$
5. The initial Distribution π where $\pi_i = P(q_1 = S_i) \ \forall \ 1 \leq i \leq N$

Given the appropriate value of N, M, A, B and π, HMM can be used as a generator to give an observation sequence

$$O = O_1 \ O_2 \ O_3 \ \ldots O_T$$

Hidden Markov Model

Evaluation Problem

Given the observation sequence $O = O_1\ O_2\ O_3\ \ldots O_T$, and model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of observation sequence given the model?

Hidden State Determination (Decoding) Problem

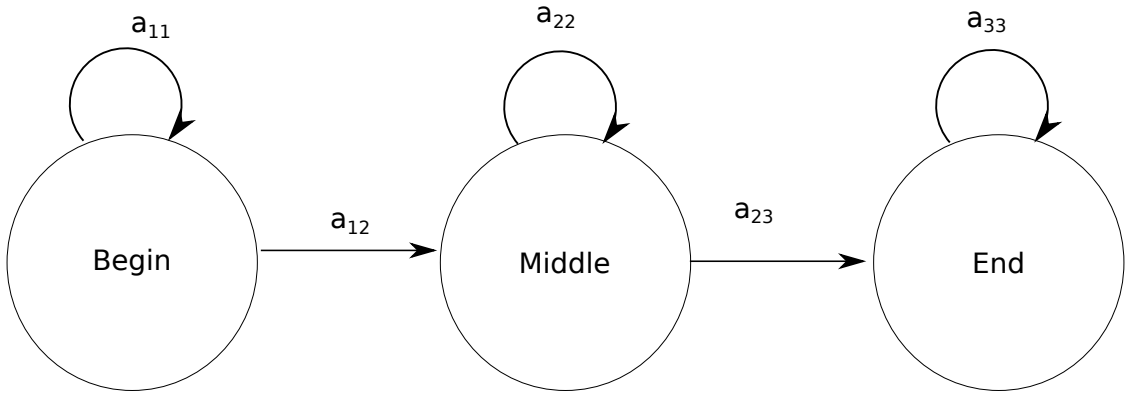
Given the observation sequence $O = O_1\ O_2\ O_3\ \ldots O_T$, and model $\lambda = (A, B, \pi)$, how do we choose corresponding state sequence $Q = q_1, q_2, \ldots, q_t$ which is optimal in some meaningful sense.

Learning Problem

How do we adjust the model parameter $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$. The observation sequence used here are called training sequence since it is used for training HMM.

Hidden Markov Model

In speech recognition, the basic unit of sound is phoneme. A phoneme is a minimal unit that serves to distinguish between meanings of words.
Due to slowly timed varying nature of the speech signal short-term spectral analysis is the most common ways to characterize speech signal. When examined over a sufficiently short period of time(between 10 and 25 msec), its characteristics are fairly stationary. For speech recognition in HMM, we assign each basic unit(phoneme) a unique HMM. The study shows that each phoneme HMM can be represented by three states, i.e. beginning, middle and end state.



Hidden Markov Model

Speech recognition consists of two main modules, feature extraction and feature matching.

- Feature Extraction
  The purpose of the feature extraction module is to convert speech waveform to some type of representation for further analysis and processing, this extracted information is known as the feature vector. Methods for Extracting feature vectors are MFCC (Mel-Frequency Cepstrum Coefficient) or LPC (Linear Predictive Coding).
- Feature Matching
  In feature matching, the extracted feature vector from an unknown voice sample is scored against an acoustic model, the model with max score wins, and its output is considered a recognized word. The acoustic model is used to score the unknown voice sample. Different types of acoustic model are GMM-Gaussian Mixture Model and VQ-Code Book.

Hidden Markov Model

Different Types of HMM used in speech recognition.

- Context-Independent Phoneme HMM
- Context-Dependent Triphone HMM
- Whole-Word HMM

Context-Independent Phoneme HMM and Context-Dependent Triphone HMM require d-state HMM for each phoneme while the Whole word HMM does not require phoneme generation. That is, the Whole word HMM assign a number of states to model a word as a whole, But requires more training data compared to other HMM.

References

[1] M. H. Moattar and M. M. Homayounpour, A simple but efficient real-time voice activity detection algorithm, in Signal Processing Conference, 2009 17th European, Aug 2009, pp. 25492553.

[2] S.nandhini and A.shenbagavalli, Voiced/Unvoiced Detection using Short Term Processing. IJCA Proceedings on International Conference on Innovations in Information, Embedded and Communication Systems ICIIECS(2):39-43, November 2014.

[3] https://www.cse.iitb.ac.in/ nirav06/i/HMM_Report.pdf

Thank You