A Literature Survey of Speech Recognition and Hidden Markov Models

Neann Mathai
University of Cape Town
Computer Science Department
MHTNEA001@uct.ac.za

Abstract

“This article highlights Hidden Markov Models (HMM) and their use in Automatic Speech Recognition (ASR) Systems. Hidden Markov Models is the most widely used speech recognition technique. Speech recognition is a field that involves a lot of different disciplines. As more and more research is being done in this field, the application possibilities of speech recognition increases, this includes military uses as well as health care usage.

This paper presents a high level view description of the basic speech recognition categories. It then goes on to discuss Hidden Markov Models and its use in ASR. Finally the paper looks briefly at some ways in which Speech Recognition Technology may be applied.”

Keywords--- Automatic Speech Recognition (ASR) Hidden Markov Models

1. Introduction

Automatic speech recognition (ASR) has been researched for more than four decades [5]. In spite of the extensive amount of time spent in this research area the ultimate goal of having speech recognition done by machine is still far from reality [1, 5]. One of the reasons that this research area has proven to be so difficult is the fact that is interdisciplinary in nature, that is, the following disciplines have to be applied [5]:

- Signal Processing
- Physics: acoustics
- Pattern Recognition
- Communication and information theory
- Linguists
- Physiology
- Computer Science
- Psychology

Real world processes generate observable signals. In the case of speech these signals are continuous, and non-stationary (they vary with time) [6,7].

When a person speaks, he uses his articulatory apparatus that consists of the lips, jaw, tongue, and velum, as can be seen in figure 1 [3]. The articulatory apparatus is used to modulate the air pressure and flow which produces a sequence of sounds [3].

![Figure 1: The Articulatory Apparatus](image)

In order to perform digital processing on an analogue speech signal, discrete time sampling and quantization of the wave form must be done [3]. Normally sampling is done at a rate of 8-20 kilohertz (kHz). The amplitude of each wave sample is then represented by one of $2^{16}$ values which is a 16-bit quantization of the unique time signal[3].

This paper will focus on Hidden Markov Models for the following reasons:

- The HMM has a mathematical structure that can be studied and analyzed.
- A HMM is relatively easy to train from a given set of training data (this is also known as an observation sequence) [3].
- HMM for speech signals are relatively easy to implement [3].
- Due to the above three mentioned points HMM have become the norm in speech recognition research.

2. ASR Approaches

2.1. Overview

There are three approaches to ASR that are as follows [5]:

- Acoustic-phonetic approach
- Statistical pattern recognition
- Artificial intelligence approach

The basic acoustic-phonetic approach is that the machine decodes the speech signal in a systematic manner by mapping the observed acoustic features of the signal to their known acoustic features and phonetic symbols. This approach is based on the theory of acoustic phonetics which suggests that in spoken language there are finite and
distinctive phonetic units. These phonetic units are each characterized by a set of properties that can be seen in the speech signal.

Pattern recognition techniques, such as the Hidden Markov Model technique, are the most popular of the speech recognition techniques. \[1,12\]. The reasons for this popularity of pattern recognition are as follows \[5\]:

- Simplicity of use
- Robustness and invariance to different speech vocabularies, users, feature sets, pattern comparison algorithms and decision rules.
- Proven high performance

The artificial intelligence approach is a combination of the acoustic phonetic approach and the pattern recognition approach. This approach automates the recognition procedure according to the way in which people applies their intelligence in analysis, and then makes a decision on the different measured feature.

Artificial Neural Networks (NNs) may also be used as a means of speech recognition. This may be thought of as a separate approach, or it may fit into any of the above three approaches. NNs is a model that attempts to mimic the human brain \[4\]. NNs very versatile and has solved problems in various fields such as computer vision, process control, and medical diagnostics. Unfortunately, NNs has a long training time. A single NN consists of many neurons. These neurons are also called perceptrons or nodes. Each neuron has a set of inputs; it then performs a computation and finally produces a single output. The neurons are connected to each other by the means of weighted paths.

The rest of this literature survey will discuss Hidden Markov Models and their use in speech recognition.

### 2.2. Hidden Markov Models

#### 2.2.1 What is a Hidden Markov Model and how is it Structured?

Hidden Markov Models are used due to the difference in speech sounds. Some sounds are sustained, such as a vowel, while other sounds are ephemeral, such as consonants. HMM are therefore ideal for this situation as they are probabilistic and can thus precisely represent the stationary and transient properties of the sound signal.

The definition of a HMM is \[7\]: A HMM is a stochastic process that cannot be observed itself (it is hidden), but the process is monitored by another set of processes that produce a sequence of observed symbols.

Generally, a HMM is made up of states, transitions (from one state to the next), and observations. It is a form of a finite state machine.

Once a model has been formulated three problems have to be solved \[1,3,5,6,7,16\]:

1. Evaluation:
   - There is a model \( \lambda \) that can be used.
   - There is a testing observation sequence \( O_t \).
   - Find \( P(O_t|\lambda) \) this is the probability that \( O_t \) was produced by this particular model. Solving this problem allows one to choose among competing models.
   - The forward-backward algorithm is used to solve this \[7\].

2. Decoding:
   - There is a model \( \lambda \) that can be used.
   - There is a testing observation sequence \( O_t \).
   - Find the most likely state path given \( \lambda \) and \( O_t \).
   - This is denoted as \( Q = q_1, ..., q_t \).
   - This problem is solved using an optimality criterion
   - The Viterbi algorithm is used to solve this \[7\].

3. Training:
   - There is a model \( \lambda \) that can be used.
   - There is a testing observation sequence \( O_t \).
   - How can the model parameters \( \lambda \) be adjusted so as to maximize \( P(O_t|\lambda) \) ?
   - The Baum-Welch algorithm is used to solve this \[7\].

An illustration of the general HMM that has been spoken of can be seen in figure 2 \[7\]. In this type of a model, any state can be revisited and the revisits do not need to take place at specific time intervals.

![Figure 2: A General Hidden Markov Model](image1)

Alternatively a non-ergodic model is shown in figure 3 \[7\]. This model is known as a left-to-right or Bakis \[6\] model which impose a temporal order to the HMM as you can only move from left to right.

![Figure 3: A non-ergodic HMM](image2)
2.2.2 HMMs as applied to Single Word Recognition

An example of how isolated word recognition can be performed is as follows [7]:

- A vocabulary of V words needs to be recognized.
- Each word and a training set of L tokens
- Using the L tokens for one of each of the V words, estimate the optimum parameters for each word and build a HMM for each word.
- For a single test word there is now an observation sequence O. Now the probability \( P(V) = P(O|\lambda V) \) for all of the models is calculated.
- The word whose model probability is the highest ie:

\[
V^* = \underset{1 \leq v \leq V}{\arg \max} [P_V]
\]

If the starting and the ending of the utterance of the word is known then it is best to use a left-to-right HMM [7]. However, if one wanted to model a long conversational speech signal, this might prove to be unfeasible [7]. A straight concatenation of the above word model may or may not prove to be successful for continuous speech recognition, and it being pursued [7].

2.2.3 Overall Recognition Process

The overall speech recognition process involves the following [6, 9]:

1. Feature Extraction
   - This is the process by which key features are extracted from the testing signals to produce the observation sets. This is done by performing a spectral and/or temporal analysis [6]. Feature Extraction involves the following:
     - Word Boundary Detection:
       When the energy on a frame exceeds a certain threshold the point is marked as the beginning of speech. The opposite is done to find the end of speech. This is known as the Energy Threshold comparison method, and the above is performed to isolate the word utterance from the starting and trailing noise [9].
     - Pre emphasis:
       The digitized speech signal is spectrally flattened by through a first order digital network [6, 9].
     - Frame Blocking:
       The speech is then blocked into frames using Hamming windows [9].
     - Cepstral Analysis:
       A vector of LPC coefficients is computed for each frame.
     - Parameter Weighing:
       This is a standard technique to weight the cepstral coefficients by a tapered window \( W_c(m) \).
     - Delta Cepstrum:
       \( c_t(m) \) is improved to include information on the temporal cepstral derivative. This is done by approximation using a polynomial which is fit over a finite window length [6, 9]. It is computed as [6]:

\[
\Delta c_t(m) = \left[ \sum_{k=-K}^{K} c_{t-k}(m) \right] \cdot G,
\]

Where:
- \( G \) = a gain term
- \( 1 \leq m \leq Q \)

At this point each word will be composed of 12 cepstral vectors and 12 delta cepstral vectors.

2. Vector Quantization [6, 9]:
   - The feature extraction produces 24 vectors that are dimensional and continuous [6, 9]. This cannot be used by a HMM [6]. Thus these vectors need to be quantized and each continuous vector is mapped into a discrete codebook index [6]. Once a codebook has been obtained using training set of vectors, all input continuous speech vectors can then be assigned the index of the “nearest” codebook vector. It should be noted that “nearest” is used in a spectral distance sense [6].

3. Hidden Markov Models:
   - Hidden Markov Models are then used to recognize a speech input based on models created using the features found from the training set.

   A diagram of the overall process is shown below in figure 4 [9].

   ![Figure 4: The Overall Speech Recognition Process](image-url)
2.2.2 Examples of Speech Recognition Systems that use HMM

- The SPHINX-II Speech Recognition System
  SPHINX was the first accurate large vocabulary, continuous, speaker-independent speech recognition system [15].

  For feature extraction SPHINX-II uses four codebooks [15]. Each codebook contains 256 entries and the four codebooks are as follows [15]:
  1. 12 LPC cepstrum coefficients
  2. 12 40-msec and 12 80-msec differenced LPC cepstrum coefficients.
  3. 12 second-order differenced cepstrum
  4. Power, 40-msec differenced power, and second order differenced power.

  The SPHINX-II system also normalizes the data from different speakers to a golden speaker cluster [15]. The golden speaker cluster is the cluster that has the maximum number of speakers. This was done using neural networks [15]. Two golden clusters were created, one male and one female cluster. Other smaller clusters were mapped to the golden speaker clusters using the created codeword-dependent neural network.

  SPHINX-II does a lot of parameter sharing when building models and uses semi-continuous hidden Markov models, and senones. SCHMMs incorporate quantization accuracy into the HMM. They estimate the discrete output probabilities by considering different codeword candidates in the vector quantization. The use of SCHMM also requires less training data when compared to discrete HMM and thus for a given data set more models can be produced. A senone is a state in a phonetic HMM. Senones are constructed by clustering the state-dependent output distributions from the different phonetic models. Senones allow for better parameter sharing and improved pronunciation optimization. Figure 5 shows a summary of the SPHINX-II system.

- The HTK Tied-State Continuous Speech Recognizer
  Like the SPHINXX-II system, HTK has a parameter generalization mechanism [14]. This allows for a balance between model complexity and parameter estimation accuracy to be obtained. Tied state refers to the fact that state distributions for corresponding states in allophone of the same phone are bunched together. This ensures that there is enough training data across all distributions. This system was trained in a number of stages, and optimized at each stage. Again, like the SPHIN-II system, HTK also used gender modeling and created different HMM for male and female voices.

3. Applications for Speech Recognition

There are numerous application for ASR, and a few military and health applications will be discussed here briefly.

1. Military Uses [13]
   Command and Control on the Move (C2OTM) is an American army project that aims to keep command and control entities mobile along with mobile troops in a war zone. Figure 6 [13] shows some of these mobile force elements that require C2OTM.

![Figure 6: C2OTM force elements and example communication of human-machine communication by voice](image)

One example of how speech recognition will be used in this project is: the foot soldier’s voice translation of what is being observed can be used to assess the battlefield situation information, and aid in weapons system selection. Another instance of voice recognition in this application is: in field repair and maintenance can be aided by a voice access to information and a helmet mounted display to show the information.

The American Navy Personnel Research and Development Center has proposed creating a combat team tactical training application. This is illustrated in
The aim of this project is to have personnel respond to ongoing combat simulations using voice, typing, trackballs, and other modes so as to communicate with both machine and with each other.

Figure 7 [13]. Combat team tactical training system concept and applications of speech based technology

An air force application being investigated by the United Kingdom’s Defense Research Agency is an application that will recognize pilots’ voices and allow them to enter reconnaissance reports. A simpler application that is being researched in terms of the air force is to allow for voice control of radio frequencies, displays and gauges in order to increase mission effectiveness and safety of the pilots.

Figure 8 [13], shows a matrix the classes of different voice applications with the interest of various military and government end users.

2. Health Care Uses
There are many, many more applications of speech recognition in society outside of the military. One such application pertaining to the health care industry is the use of speech recognition in automatic medical transcription [8,11]. This avenue is being seriously considered as it may prove to be more cost effective with a projected savings of $230 a week in 1998.

Conclusions
This paper has looked at speech recognition and how this can be solved using Hidden Markov Models. HMM are widely used for speech recognition as they are relatively easy to use and they are mathematically backed and as a result they can be analyzed.

The overall speech recognition process involves extracting features from the signal, performing a vector quantization on these features and then using the HMMs to recognize a particular utterance, based on probabilities.

The paper finally touched on a few applications areas that are open to speech recognition as systems keep getting better.

References


