Speech Recognition using Hidden Markov Model

Mrs. Varsha Degaonkar¹ Dr. Anju V. Kulkarni² Dr. Radhika Menon³
¹Research Scholar, E&TC Dept., JSPM's RajarshiShahu College of Engg, Tathawade, Pune
², ³Professor, PadmashreeDr. D. Y. Patil Institute of Engg & Technology, Pimpri, Pune
varshadegaonkar@gmail.com¹, anju_k64@yahoo.co.in², radhika_tharoor@gmail.com³

Abstract. Speech technology and systems in human computer interaction have witnessed a steady and important advancement over last two decades. Today, speech technologies are commercially available for boundless but interesting range of tasks. These technologies permit machines to respond correctly and consistently to human voices, and provide useful and valuable services. In the present era, mainly Hidden Markov Model (HMMs) based speech recognizers are used. This paper aims to present a speech recognition system using Hidden Markov Model. Hidden Markov Model Toolkit (HTK) is used to develop the system. It is used to recognize the isolated words using acoustic word model.

Keywords: Automatic Speech Recognition (ASR), HMM model, Hidden Markov Model Toolkit (HTK), Human Machine Interaction.

I. INTRODUCTION

Speech is the most natural way of communication. It provides an efficient means of man-machine communication. Generally, transfer of information between human and machine is accomplished via keyboard, mouse etc. But human can speak more quickly instead of typing. Speech input offers high bandwidth information and relative ease of use [10]. It also permits the user’s hands and eyes to be busy with a task, which is particularly valuable when users are in motion or in natural field settings. Speech recognition can be defined as the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. It is done by means of Algorithm implemented as a computer program. Speech processing is one of the exciting areas of signal processing. Based on major advanced in statically modeling of speech, automatic speech recognition today find widespread application in task that require human machine interface such as automatic call processing [3].

In this paper, we present a speech recognition system for isolated words. Hidden Markov Model (HMM) is used to train and recognize the speech and MFCC is used to extract the features from the speech-utterances. To carry out this, Hidden Markov Model toolkit (HTK) [2] designed for speech recognition is used. HTK is developed in 1989 by Steve Young at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED). Initially, HTK training tools are used to train HMMs using training utterances from a speech corpus. Then, HTK recognition tools are used to transcribe unknown utterances and to evaluate system performance by comparing them to reference utterances.

© Elsevier, 2013
Apart from introduction in section 1, the paper is organized as follows. Section 2 presents the architecture and functioning of proposed ASR. Section 3 describes the Hidden Markov Models and HTK. Section 4 deals with implementation of Speech Recognition using Hidden Markov Model. Section 5 concludes the paper.

II. SPEECH RECOGNITION SYSTEM ARCHITECTURE

The speech recognition system architecture is shown in fig. 1. It consists of two modules, training and testing module. Training module generates the system model which is used during testing. The various phases used during ASR are:

**Preprocessing:** Speech-signal is an analog waveform which cannot be directly processed by digital systems. Thus preprocessing is necessary to transform the input speech into a form that can be processed by recognizer. To accomplish this, the speech-input is digitized or sampled. The sampled speech-signal is then processed through the first-order filters to spectrally flatten the signal. This process, known as pre-emphasis, increases the magnitude of higher frequencies with respect to the magnitude of lower frequencies. The next step is to divide the speech-signal into the frames with frame size ranging from 10 to 25 milliseconds and an overlap of 50%−70% between consecutive frames [10].

![Speech Recognition System Architecture](image)

**Feature Extraction:** The goal of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that can somehow be computed or estimated through processing of the signal waveform that is speech waveform. Such parameters are termed as features. The feature extraction process is expected to discard irrelevant information to the task (ex. Silence part, noise) while keeping the useful one. It includes the process of measuring some important characteristic of the signal such as energy or frequency response (i.e. signal measurement), augmenting these measurements with some perceptually meaningful derived measurements (i.e. signal parameterization), and statically conditioning these numbers to form observation vectors [10].

**Model Generation:** The model is generated using various approaches such as Hidden Markov Model (HMM), Artificial Neural Networks (ANN), Dynamic Bayesian Networks (DBN), Support Vector Machine (SVM) and hybrid methods (i.e. combination of two or more approaches). Hidden Markov model is used in some form or another in virtually every state-of-the-art speech and speaker recognition system.

**Speech Transcription:** Speech Transcription component recognizes the test samples based on the acoustic properties of word. The classification problem can be stated as finding the most probable sequence of words $W$ given the acoustic input, which is computed as:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)}$$

(1)

Given an acoustic observation sequence $O$, classifier finds the sequence $W$ of words which maximizes the probability $P(O|W)P(W)$. The quantity $P(W)$ is the prior probability of the word which is estimated by the language model. $P(O|W)$ is the observation likelihood, called as acoustic model.

III. HIDDEN MARKOV MODELS AND HIDDEN MARKOV MODEL TOOLKIT (HTK)

A. Introduction

The Hidden Markov model (HMM) is a very powerful mathematical tool for modeling time series. It provides efficient algorithms for state and parameter estimation, and it automatically performs dynamic time
warping for signals that are locally squashed and stretched. It can be used for many purposes other than acoustic modeling.

**B. Markov Chains [4]**

Hidden Markov models are based on the well-known Markov chains from probability theory that can be used to model a sequence of events in time. Fig. 2 shows such a graphical network representation of such a model, it has two states a and b and some connections indicated by arrows that show how one can get from one state to another1. The topology of the network shows an important property of Markov chains, namely that the next state only depends on the current state the model is in, regardless of how it got in the current state; this property is often referred to as the Markov property. By starting in one of the two states and at each time step moving through the model following the arrows out of the current state to the other state or once again to the same state, sequences of a’s and b’s can be generated.

![Fig. 2 A Markov Chain](image)

The arrows leaving a state are annotated with a probability that indicates how likely it is that this particular transition out of the state will be chosen. The distribution indicates how likely each state is to be the start state, in Fig.2 both states are equally likely to be the start state. Using these probabilities a Markov model can be used for recognition.

**C. Definition of Hidden Markov models**

The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states $Q$, an output alphabet (observations) $O$, transition probabilities $A$, output (emission) probabilities $B$, and initial state probabilities $\pi$. The current state is not observable. Instead, each state produces an output with a certain probability (B). Usually the states $Q$, and outputs $O$, are understood, so an HMM is said to be a triple $(A, B, \pi)$.

**Description of HMM:** For the description fig. 3 shows an example of Hidden Markov Model [5]. The model consists of a number of states, shown as the circles in figure. At time the model is in one of these states and outputs an observation (A, B, C or D). At time $t+1$ the model moves to another state or stays in the same state and emits another observation. The transition between states is probabilistic and is based on the transition probabilities between states which are given in state $j$ at time $t+1$. Notice that in this case $A$ is upper triangular. While in a general HMM transitions may occur from any state to any other state, for speech recognition applications transitions only occur from left to right i.e. the process cannot go backwards in time, effectively modeling the temporal ordering of speech sounds. Since at each time step there must always be a transition from a state to a state each row of $A$ must sum to a probability of 1. The output symbol at each time step is selected from a finite dictionary. This process is again probabilistic and is governed by the output probability matrix $B$ where $B_{jk}$ is the probability of being in state $j$ and outputting symbol $k$. Again since there must always be an output symbol at time $t$, the rows of $B$ sum to 1. Finally, the entry probability vector $\pi$, is used to described the probability of starting in described by the parameter set $\lambda = [\pi, A, B]$

![Fig. 3 A Five State Left-Right, Discrete HMM for Four Output Symbols](image)

$$
A = [a_{ij}] = \begin{bmatrix}
0.9 & 0.1 & 0 & 0 & 0 \\
0 & 0.6 & 0.4 & 0 & 0 \\
0 & 0 & 0.7 & 0.3 & 0 \\
0 & 0 & 0 & 0.4 & 0.6 \\
0 & 0 & 0 & 0 & 1.0 \\
\end{bmatrix}
$$

150
A HMM is characterized by the following:

- \( N \), the number of states in the model. The individual states are denoted as \( S = \{S_1, S_2, \ldots, S_n\} \) and the system state at time \( t \) as \( q_t \).
- \( M \), the number of distinct observation symbols per state, i.e. the discrete alphabet size. The individual symbols are denoted as \( V = \{v_1, v_2, \ldots, v_m\} \).
- The transition probability distribution \( A = \{a_{ij}\} \) where, each \( a_{ij} \) is the transition probability from state \( S_i \) to state \( S_j \). Clearly, \( a_{ij} \geq 0 \) and \( \sum_{j} a_{ij} = 1, \forall i \).
- The observation symbol probability distribution \( B = \{b_{jk}\} \) where, each \( b_{jk} \) is the observation symbol probability for symbol \( v_k \), when the system is in the state \( S_j \). Clearly, \( b_{jk} \geq 0, \forall i,j \) and \( \sum_{k} b_{jk} = 1, \forall j \).
- The initial state distribution \( \pi = \{\pi_j\} \) where, \( \pi = P[q_1 = S_j], 1 \leq j \leq N \). HMM model can be specified as \( \lambda = (A, B, \pi, M, N, V) \). Here, HMM is represented as \( \lambda = (A, B, \pi) \) and assume \( M, N \) and \( V \) to be implicit.

**D. Use of HMM in Speech Recognition**

The HMM can be used to model a unit of speech like a phoneme, a word, a sentence. In order to use HMMs in speech recognition, one should have some means to achieve the following:

**Evaluation**: Given the observation sequence \( O = (o_1, o_2, \ldots, o_T) \) and a HMM \( \lambda = (A, B, \pi) \) to choose a corresponding state sequence \( Q = q_1, q_2, \ldots, q_T \) which optimal in some meaningful sense, given the HMM.

**Training**: To adjust the HMM parameters \( \lambda = (A, B, \pi) \) to maximize \( P(O | \lambda) \).

The following are some of the assumptions in the Hidden Markov Modeling for speech [5]:

1. Successive observations (frames of speech) are independent and therefore the probability of sequence of observation \( P = (o_1, o_2, \ldots, o_T) \) can be written as a product of probabilities of individual observations, i.e.
   \[
   O = (o_1, o_2, \ldots, o_T) = \prod_{i=1}^{T} P(o_i)
   \]

2. Markov assumption: The probability of being in a state at time \( t \), depends only on the state at time \( t-1 \).

**E. Hidden Markov Model Toolkit**

HTK is a toolkit for building Hidden Markov Models (HMMs). It is an open source set of modules written in ANSI C which deal with speech recognition using the Hidden Markov Model. HTK mainly runs on the Linux platform. However, to run it on Windows, interfacing package Cygwin [8] can be used.

**IV. IMPLEMENTATION OF SPEECH RECOGNITION USING HIDDEN MARKOV MODEL**

In this section, implementation of the speech recognition system using HMM has been presented.

**A. System Description**

Speech recognition system is developed using HTK toolkit on the Windows platform. To estimate the parameters of a set of HMMs using training utterances and their associated transcriptions, Hidden Markov Model Toolkit (HTK) training tools are used. Then, unknown utterances are transcribed using the HTK recognition tools. System is trained for 50 words. To recognize the speech Word model is used.

**B. Database Generation**

Speech database is required for Training and testing of a speech recognition system. A data-set of 50 words is used by system for training it. The data is recorded using unidirectional microphones. Seven speakers have
used to record totally $420(7 \times 4 \times 15)$ speech files. The speech is recorded at 16 kHz & in .wav format using “SIL” Speech Analyzer software tool in normal lab and normal room conditions.

C. Feature Extraction
In feature extraction step, the data recorded is parameterized into a sequence of features. For that, HTK tool HCopy is used. For parameterization of the data, Mel Frequency Cepstral Coefficient (MFCC) is used. The MFCC is the extensively used method for feature extraction in speech recognition system [1]. In MFCC, the frequency bands are located logarithmically; so it approximates the human ear system response more closely than any other system. MFCC technique is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients, the speech signal is sampled at 16 kHz, and then processed at 10 ms frame rate with a Hamming window of 25 ms to minimize the discontinuities of a signal. Then DFT is used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is calculated. After the warping, 12 MFCC coefficients are obtained. Finally the Inverse Discrete Fourier Transform is used for the cepstral coefficients calculation. It transforms the log of the quefrency domain coefficients to the frequency domain where $N$ is the length of the DFT [4]. MFCC can be computed by using the formula,

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700})$$  \hspace{1cm} (2)

D. Training the HMM
For training the HMM, a prototype of HMM model is created (dictionary template), which are then re-estimated using the data from the speech files. Apart from the models of vocabulary words, model for silent must be included. For prototype models, 5-9 states of HMM are used in which the first and last are non-emitting states. The prototype models are initialized using the HTK tool HInit which initializes the HMM model based on one of the speech recordings. Then HRest is used to re-estimate the parameters of the HMM model based on the other speech recordings in the training set.

E. Performance Evaluation
During evaluation, system is responsible for generating the transcription for an unknown utterance. The model generated during the training phase is responsible for evaluation. In order to evaluate the system performance, speakers are asked to utter each word at least once a time. Overall word-accuracy and word-error rate of the system is calculated. Word error rate is a common metric of the performance of a speech recognition or machine translation system [11]. Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N}$$

Where $S$ is the number of substitutions, $D$ is the number of the deletions, $I$ is the number of the insertions, $N$ is the number of words in the reference.

V. CONCLUSIONS
In this article, we have presented speech recognition architecture and the statistical method of HMM. The HMM training process can automatically determine word and phone boundary information during training. This means that it is relatively straightforward to use large training corpora. It is the major advantage of HMM which will extremely reduce the time and complexity of recognition process for training large vocabulary.

The presented system is used for recognition of isolated words using acoustic word model. The training of the system has done using 50 words. During the development of the system, the training data has collected from the seven different speakers. The system has tested in the normal room environment. The implementation of the system has done using Hidden Markov Model Toolkit (HTK) on Windows operating system. The HMM approach along with MFCC is used for good recognition result.
REFERENCES


