# Adaptive HMM based Speech Recognition to Recognize Multi-lingual Sentence

Shrurti Gupta
ECE, DIT University
Dehradun

Kashif Shabeeb
ECE, DIT University,
Dehradun

Sonika Singh,
Ph.D.
ECE, DIT University,
Dehradun

Sandeep Sharma,
Ph.D.
ECE, DIT University
Dehradun

## ABSTRACT
Hidden Markov Models (HMMs) provides an effective framework for the modeling of time-varying sequence of spectral vector. An approach of mapping signal to discrete signal is to define it as a set of acoustic featured symbol over a minimal but constant time interval. The aim of proposing this paper is to recognize the speech sample using hidden markov model (HMM) with the use of cepstrum feature of our given speech sample within an adaptive interval of time for which pitch period is determined and divides the sample in accordance with this period. Secondly the phonetics or exact "pronunciation" of the word needs to be defined. These are established by associated rule probability where probability is done on word's pronunciation.

## General Terms
Multilingual Speech Recognition

## Keywords
Hidden Markov Model (HMM), Associated Rule Probability, and Pitch Detection Algorithm.

## 1. INTRODUCTION
Earlier adaptive framing [1] was used for the purpose of speech recognition in English language. Recently interests have been shown in the development of multilingual form of recognition of speech. Hidden Markov Model provides efficient recognition over large vocabulary for continuous speech recognition [2] to speaker dependent as well as speaker independent signals. In this paper two main models of recognition; feature extraction [3] and feature matching have been described. The most important segmentation of constant time framing [4] for increasing efficiency on the cost of phoneme length and more importantly complexity of the phoneme (an important feature of multilingual communications) have been used in our work. Windowing'scharacteristic performs a function of smoothing the varying estimates and reduces the copping negative effects [5], [6] which is also used. Pitch period, a very important parameter used in the synthesis of speech signal,linguistic and phonetics [7] is determined using this feature of speech. Here in this adaptive HMM, pitch period framing is used**.**

## 1.1 Hidden Markov Model
Speech recognition can be classified as –

A. Single word recognizer B. Continuous word recognizer

C Speaker dependent D. Speaker independent

HMM: creates stochastic models from unknown utterances and compares the problem that the unknown utterance was generated by each model. Here each state transition will emitter absorbs the observation in accordance to some PDF. The sequence cannot be uniquely defined from the sequence of observation and is hidden therefore, Hidden Markov Model HMM [8] is defined by a set of 3 states observation symbols and the 3 matrices.

$$M = [\pi, A, B] \quad (1)$$

Where

$\pi=\pi_i$ initial state probability

$A=a_{i,j}$ state transition probability

$B=b_{i,j,k}$ symbol emission probability

## 1.2 Basic algorithms
A) Classification of an unknown observation (recognition) sequence.
B) Training models from a set of training data.
C) Evaluation of problem of an observation sequence.

## 2. PROPOSED WORK
## 2.1Classification using HMM's
HMM's use a maximum aposteriori likelihood classification (recognition) rule

$$\text{chosen class} = \text{argmax class } [P(M_{class}|0] \quad (2)$$

Aposteriori likelihood is given as;

$$I[M_{class}|o] = P\{O|M_{class}) P(\text{class}) \quad (3)$$

Evaluation of P (o|m) is also called DECODE which is characterized as Forward Decoder and Backward Decoder. Former is the sum of probabilities on Exit State while later starts at exit state and applies the observation in reverse order.

Forward Decoder

$$P(O|M) = \sum_i \varepsilon_{(terminal\ \ state)}\, \alpha_{T+1}(i) \quad (4)$$

Backward Decoder:

$$P(O|M) = \sum_{i=n}^{N} \pi_i \beta_t(i) \quad (5)$$

Thus resultant

$$P(O|M) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad 1 < t < T+1 \quad (6)$$

*Programming counterparts*
1) Numerical underflow: Most of computers under flow as numbers after multiplication of several probabilities become very small ($=10^{-2000}$).Thus log format/scaling must be entertained for better performance.

2) Numbers of free variables for training data must be reduced for training data by tying states for more complexity.
3) Missing Observation Problem: It occurs because of non-trained symbols by which their probability becomes Zero. This is improved by applying a lower limit on power density function of observed value.

## 2.2 The Baum–Welch algorithm

This algorithm estimates the parameters of HMM. The o/p object contains additional parameters as they are replaced by estimated values.

For generation of HMM (λ) we need

1. N States of model, S= {S1,S2,S3,….$S_n$}
2. Possible output symbols,

$$\boldsymbol{M} = \sum\{\boldsymbol{\sigma 1}, \boldsymbol{\sigma 2}, \dots\}$$

3. State transition Probability distance A= {$a_{ij}$}

   $a_{ij}$ is the probability state at t+1 → $S_j$ given state at t is S

4. Observation symbol probability distribution B={$b_i(\sigma_k)$},

$b_i(\sigma_k)$→ probability ($\sigma_k$) emitted in state $S_j$.

5. Initial state distribution $\pi = \pi_j$

$\pi_j$ → Probability that model is in state $S_j$ at t=0.

Baum Welch Algorithm improves HMM by using expectation minimization. Let $X_t$ be a discrete hidden random variable with possible values. We assume that P ($X_t|X_t$-1) is independent of time T, and this leads to Time Independent Stochastic Transition Matrix's definition.

$$A = \{a_{ij}\} = P\ (X_{t=j}|X_{t-1}=i). \quad (7)$$

Initial state distribution (i.e. when **t=1**) is given by

$$\boldsymbol{\pi_i}=\mathbf{P(X_{1=i})}. \quad (8)$$

The observation variables $\mathbf{Y_t}$ can take one of **K** possible values. The probability of the observation at time **T** for state **j** is given by

$$\mathbf{B_j(y_t)=P(Y_t= y_t|X_t=j)}. \quad (9)$$

Taking into account all the possible values of $Y_t$ and $X_t$ we obtain the **K** by **N** matrix

$$\mathbf{B = \{(b_j y_i)\}}.$$

An observation sequence is given by

$$\mathbf{Y= (Y_1=y_i, Y_2=y_2,\dots,Y_t=y_t).}$$

Thus hidden Markov chain can be described by

$$\theta = (A, B, \pi). \quad (10)$$

The Baum–Welch algorithm finds a local maximum for

$$\theta^* = \max P(Y|Q) \quad (11)$$

(i.e. the HMM parameters $\theta$ that maximize the probability of the observation.)

*1.1.1 Algorithm:*

Set $\theta = (A, B, \pi)$ with random initial conditions. If the initial information of the Parameters is available then also they can be set directly.

***Forward procedure***

Let $\alpha_i(t) = P(Y_1, \dots, Y_{t=y_t}, X_{t=i})$ the probability of seeing the $\mathbf{y_1, y_2, y_3, \dots}$ and being in state **i** at time **t**. This is found recursively:

1. $\boldsymbol{\alpha_i(1)}$
2. $\alpha_j(t+1) = b_j(y_{t+1}) \sum_{i=1}^{N} \alpha_i(t) a_{ij}$

***Backward procedure***

Let $\beta_i(t) = P(Y_{t+1}, \dots, Y_{T=y_t}|X_{t=i}, \theta)$ that is the probability of the ending partial sequence $y_{t+1},\dots,y_T$ given starting state *i* at time *t*. We calculate $\beta_i(t)$ as,

1. $\beta_i(t) = 1$
2. $\beta_i(t) = \sum_{i=1}^{n} \beta_j\ (t+1) a_{ij}\, b_j(y_t)$

*Update*

We can now calculate the temporary variables:

$$\gamma_i(t) = P(X_{t=i}|\gamma, \theta) \quad (12)$$

which is the probability of being in state **i** at time **t** given the observed sequence **Y** and the parameters θ

$$\xi_{ij}(t) = P(X_{t=i}, X_{t+1=j}|\gamma, \theta)$$

$$= \frac{\alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}{\sum_{k=1}^{N}\sum_{t=1}^{N} \alpha_k(t) a_{kl} \beta_j(t+1) b_l(y+1)}$$

$$= \frac{\alpha_i(t) a_{ij} \beta_j(t+1) b_j(y_{t+1})}{\sum_{k=1}^{N} \alpha_k(t) \beta_k(t)}$$

which is the probability of being in state **i** and **j** at times **t** and **t+1** respectively given the observed sequence **Y** and parametersθ.

θ can now be updated:

$$\boldsymbol{\pi_i^* = \gamma_i(1)}$$

which is the expected frequency spent in state **i** at time **1**

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \quad (14)$$

Here this is the predicted number of transitions from one state I to another state j in contrast to the expected over all number of transitions away from state i. To clarify this, it does not mean transitions of one state I to another state j, but to any of the state including I itself. This equates to the total number of times state I is observed one after the other from t=1 to t=T-1.

$$b_i^* = \frac{\sum_{t=1}^{T-1} ly_t=v_k=\gamma_i(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \quad (15)$$

where

$$ly_t = v_k = \begin{cases} 1, if\ y_t = v_k \\ 0, otherwise \end{cases} \quad (16)$$

is a function indicator and is $\mathbf{b_i^*}$, the so predicted number of times the Output observations is equal to $\mathbf{v_k}$ while when in state **i** over the predicted number of times is in state **i**.

The steps above are now continually repeated till a desired level of convergence is observed.

**Note:** It is possible to over-fit a particular data set. That is $\mathbf{P(Y|\theta_{final}) > P(y|\theta_{true})}$.The algorithm also does not guarantee a global maximum.

## 2.3 Pitch Detection Algorithm in Speech

These algorithms are categorized into

a) Time Domain Based Tracking

b) Frequency Domain Based Tracking

c) Joint Time Frequency Based Tracking

d)

***Some of the most used algorithms are gives as***

1. Short Term Autocorrelation Function Method(ACF)
2. Harmonic Product Spectrum(HPS)
3. Robust Algorithm for Pitch Tracking(RAPT)
4. Average Magnitude Differenc
5. e Function(AMDF)
6. Simple Inverse Filtering Tracking(SIFT)
7. Direct Frequency Estimation(DFE)
8. Cepstrum Pitch Determination(CPD)

Some of the modifications are done to enhance the performance of the system. Some methods are described as;

***(A) Modified Autocorrelation Function Method (MACF)*** uses center clipping technique for flattening of spectrum signal in Preprocessing Stage .relationship between input signal and center clipped signal is

$$y(n) = clc[x(n)] = \begin{cases} (x(n) - C), x(n) \geq C \\ 0, |x(n)| < C \\ (x(n) + C, x(n) \leq -C \end{cases}$$

where $C_L$ is the Clipping threshold which is about 50% if the maximum absolute signal. Autocorrelation function is also periodic with same period.

$$R(m) = \frac{1}{N}\sum_{n=0}^{N-1-m} x(n).x(n+m), 0 \leq m < M_o \quad (17)$$

Maximum R(m) results when the value of Pitch is equal to the value of 'm'.

***(B)Normalized Cross Correlation Function Method(NCCF)*** This method follows variations in pitch and amplitude of speech.

$$NCCF(m) = \frac{\sum_{n=0}^{N-m-1} x(n).x(n+m), \quad 0 \leq m \leq M}{\sqrt{\sum_{n=0}^{N-1-m} x^2(n).\sum_{n=0}^{N-m-1} x^2(n+m)}} \quad (18)$$

Where N=length of analyzed frame, 'm'=lag, $M_o$=no of autocorrelation points to be computed.Here voiced and unvoiced decision is done on the frames energy basis. For voiced speech lag corresponding to highest peak of NCCF is considered as Pitch Period

***(C) Average Difference Function Method (AMDF):***
Here a signal is formed by difference between original and delayed speech and at each delay value absolute magnitude is taken.

$$Dx(m) = \frac{1}{N}\sum_{n=0}^{N-1-m} |x(n) - x(n+m)|, \quad 0 \leq m \leq M \quad (19)$$

Where x (n) is the samples of analyzed speech, x(n-m) is the samples time shifted on M samples, N is the frame length. Here voiced/unvoiced decision is computed on amount of frame energy.

***(D) Cepstrum Pitch Determination(CPD):***
Here the cepstrum of voiced speech has strong peak and is more useful than other pitch detecting algorithms. Sequence of voiced speech is represented as

$S(n) = e(n) * h(n)$

Where e (n) is the source extraction sequence and h(n) is the vocal tract impulse response.In frequency domain

$$S(\omega) = E(\omega).H(\omega)$$

where $S(\omega)=F\{s(n)\}$, $E(\omega)=F\{e(n)\}$ and $H(\omega)=F\{h(n)\}$.

Symbol *F* stands for Discrete Fourier Transform (DFT)

Cepstrum is bifurcated and the part of cepstrum which represents Source signal is extracted and then the Pitch Period is evaluated.

The real Cepstrum of discrete signal S (n) is given as:

$$C(m) = \frac{1}{N}\left\|\sum_{k=0}^{N-1} s(k).e^{-j\frac{2\pi}{N}mk}\right\| \quad (20)$$

Here, S (k) is magnitude of Logarithmic Spectrum of s(n)

$$S(k) = log\left\|\sum_{n=0}^{N-1} S(n).e^{-j\frac{2\pi}{N}nk}\right\| \quad (21)$$

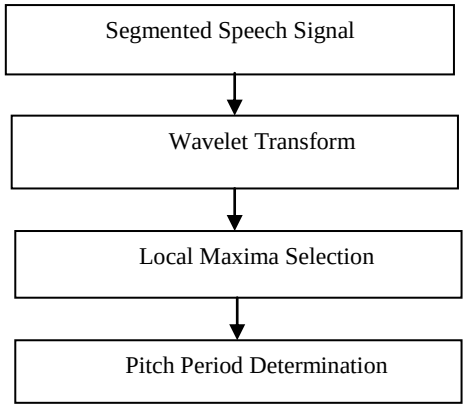
If the value of peak exceeds a particular threshold value then the speech is voiced and location of peak is the pitch period and if it is below the threshold value then it is considered as unvoiced speech to Zero Crossing Count.

Accuracy of PDA is measured by:

1. Classification Error (CE): It is the % of unvoiced frames that are distinguished as voiced and voiced frames are termed as unvoiced.

2. Gross Error (GE): % of voiced frames with respect to an estimated frequency or fundamental frequency quantity that diverges from the some reference quantity more than 20%.

A Dyadic Discrete Wavelet Transform divides the segmented speech into several bands.




The peaks in the spectra cause spectral shaping which needs to be removed by Spectral Flattening. Usually two techniques are used for Spectrum Flattening

1) A Filter Bank Method
2) Centre Clipping Method
3) The autocorrelation of a sequence is correlation of a sequence with itself, the autocorrelation of a sequence x(n) is defined as
4) $\emptyset_x(m) = \lim_{N\to\infty} \frac{1}{2N+1}\sum_{n=-N}^{N} x(n).x(n+m).$ (22)

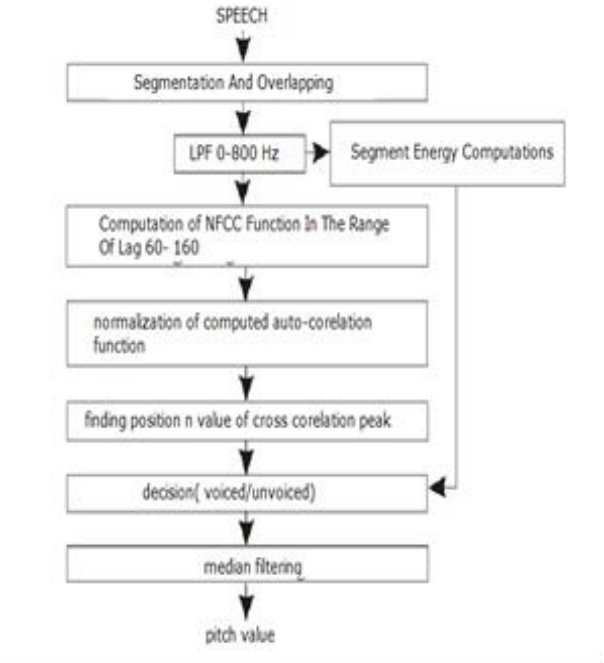



**Fig.2.1 NCCF Pitch Detector**

Some symbols which are used in the algorithm:

*N*: frame size in sample.

*s*[*n*]: the voiced speech of the m[th] frame, where $0 \leq n < N$.

$SF_m[k]$: the frequency response of *s[n]*, where $0 \leq k < N$.

$YF_m[k]$: the pass band frequency response $SF_m[k]$,

where$0 \leq k < N$.

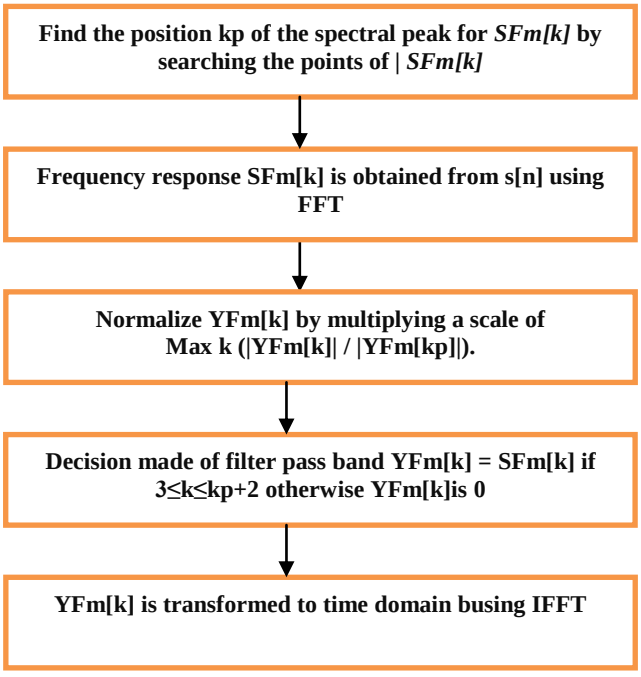



**Fig.2.2. Pitch detection algorithm based on filter bank**

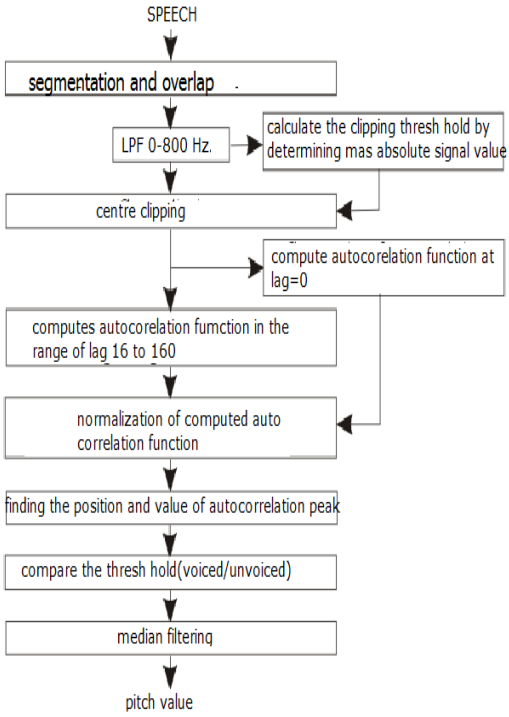



**Fig.2.3MACF Pitch Detector**

## 3. RESULTS

**Table 3.1 shows terEsults obtained for various languages**

| Language | Samples | Data | Accuracy (rounded off) |
|---|---|---|---|
| English | 40 | Alphabets | 100% |
| Hindi | 40 | Varnamala (aphabets) | 99% |
| Urdu | 40 | Alphabets | 94% |
| English | 20 | Word | 98% |
| Hindi | 20 | Word | 95% |
| Urdu | 20 | Word | 97% |
| English | 20 | Continuous | 92% |
| Hindi | 20 | Continuous | 96% |
| Urdu | 20 | Continuous | 98% |

## 4. CONCLUSION

Here in this paper the hidden markov model for the recognition of speech by detecting the adaptive pitch of the speech sample is proposed. It is concluded that with the use of different training algorithms and recognition algorithms we can recognize any language speech. Efficient recognition of large vocabulary continuous speech recognition is provided to both types of speakers (i.e. Speaker dependent as well as speaker independent) and reduces the complexity of phoneme.

Multilingual speech can be recognized using the above technique and its phonemes can be tracked using pitch detection technique. Future work can be done on text to speech conversion for different languages. This could help drivers to understand road side boards while driving through different states.

## 5. ACKNOWLEDGEMENT

I feel expedient to express my profound indebtedness and deep sense of gratitude and sincere thank to Dr. Sandeep Sharma Head of Department of electronics and communication engineering and Dr. Sonika Singh , Dehradun Institute of Technology, Dehradun, for their valuable guidance, inspiration, encouragement and whole hearted cooperation in understanding the relevant field and help in presenting this work. Their keen interest and efforts in planning the work in this form cannot be expressed in words as they devoted their valuable time in discussion and critical analysis of the work. I would also like to thank the other member of the department, for taking the time to review my research.

## 6. REFERENCES

[1] Yeh-HuannGoh, ParamesraNRaveendran "HMM-Based Speech Recognition Using Adaptive Framing" IEEE TENCON 2009.

[2] Gauvain, J. ; Lamel, L. "Large-vocabulary continuous speech recognition: advances and applications" Published in: Proceedings of the IEEE (volume:8) in Aug. 2000

[3] Banik, M.Eity, Q.N. ; Lisa, N.J. ; Hassan, F"Japanese phonetic feature extraction for automatic speech recognition" published in: signal and image processing (ICSIP), 2010 international conference on Date 15-17 Dec. 2010

[4] S. Young, "large vocabulary continues speech recognition: a review," IEEE signal processing magazine. vol. 13(5), PP. 45-57, 1996

[5] Lawrence R. Rabiner, "a Tutorial on Hidden Markov Models and selected applications in Speech Recognition," proceeding of the IEEE, vol. 77, No. 2, february 1989.

[6] Joseph W. Picone, "Signal modeling techniques in speech recognition,"proceeding of the IEEE, vol. 81, no. 9, September 1993.

[7] Kadambe, S., Boudreaux-Bartels, G.F., "application of the Wavelet Transform for Pitch Detection of Speech Signals", IEEE transactions On information theory, volume 38, issue 2, part 2, March 1992 Page(s): 917-924

[8] Rabiner, Lawrence. "First Hand: The Hidden Markov Model".IEEE global history network.retrieved 2 October 2013.

[9] J.ABilmes "Gentle Tutorial on EM Algorithms and its application to Parameter Estimation for Gaussian mixture and hidden Markov Models.

[10] Ibrahim M. El-Henawy,et. al., "Recognition of phonetic Arabic figures via waveletbased Mel Frequency Cepstrum using HMMs", HBRCJournal,September 2013