# Issues and Limitations of HMM in Speech Processing: A Survey

Chandralika Chakraborty
Dept. of IT,
Sikkim Manipal Institute of Technology, Sikkim.

P.H. Talukdar
Dept. of IT,
Kaziranga University, Assam.

## ABSTRACT
Speech is the most natural way of communication among humans. This mode of communication is constituted of two parts, namely sound and sense. The intelligent production and synthesis of speech has intrigued man himself for long and efforts at automated speech recognition, has gone through various phases. Hidden Markov Models (HMMs) provide a simple and effective framework for modeling time-varying spectral vector sequences. Application of HMMs to speech recognition has seen considerable success and gained much popularity. As a consequence, almost all present day speech recognition systems are based on HMMs.

The current paper presents a brief study on the HMM based technique applied to speech recognition and also discusses the issues and limitations of HMMs in speech processing.

## Keywords
Speech recognition, speech representation, Hidden Markov Model, implementation Issues, limitations, challenges.

## 1. INTRODUCTION
Speech is the most natural and primary means of communication between humans. This mode of communication developed over many many years, through the evolution and associated changes in physiology, climate and society. Intelligent activity of communication is also found in other animals, which are also known to communicate with sound. However, speech based communication system as developed as man, is not known to be found in other forms of life.

The current paper is based on studies on HMM based speech recognition systems and presents the approach along with its various issues and limitations. The following sections 2& 3 introduce the mechanism of speech production and representation respectively. Section 4 introduces the Hidden

Markov Model before presenting the applications in Speech Recognition and various issues in subsequent sub-sections.

Section 5 presents the various HMM based speech recognizers while section 6 and 7 discusses on the limitations and challenges of HMMs in speech processing, with finally section 8 concluding the paper.

## 2. SPEECH PRODUCTION
Speech production in humans is a physiological activity which involves the co-ordination of the lungs, vocal cords, vocal tract, palate, tongue, teeth, lips and nasal cavity. Air enters the lungs when breathing, and when it is expelled from the lungs through the trachea it causes the vocal cords to vibrate. The vocal cords, also known as vocal folds are composed of twin infoldings of membranes stretched horizontally across the larynx, which is also called voice box [1]. The vibration causes the air flow to be converted into quasi-periodic pulses of air which then becomes sound which is the wave phenomenon that results from air pressure variations. The pulses are frequency shaped by the oral cavity and the nasal cavity [2]. The vocal folds open during inhalation, close when holding one's breath and vibrate while speaking.

The physiological parameters like length of vocal folds, vocal tract, thickness of larynx ,etc., influence the acoustical parameters like pitch, fundamental frequency, formant frequency, etc., which results in the differences between male and female voices. These acoustical parameters play a significant role in Speaker Gender Recognition. Men and women have different vocal fold sizes. The vocal folds in men are between 17.5 mm & 25 mm in length while the adult female folds are between 12.5 mm and 17.5 mm in length [2]. As adult males have larger vocal folds they are usually lower pitched than the females. Another organ, the vocal tract, is found to be larger in length in males than in females. The ratio of the total length of female vocal tract to that of the male is 0.87 [3]. The female larynx also differs from the male larynx in thickness [4]. Due to these features, the determining parameters of speech recognition vary between the male voice and the female voice. The pitch of the female voice ranges between 170 Hz – 275 Hz [5] while the male voice is approximately one octave lower [6] at about 112Hz – 116 Hz [7].

The fundamental frequency for an adult male and an adult female is around 131 Hz and 220 Hz respectively [8]. Formant frequency of females is said to be scaled upward in frequency by about 20% compared to the average male formant pattern [9]. Cited literature [10] also reported that male formants were lower in frequency than female formants.

## 3. REPRESENTATION OF SPEECH
Speech signal is a slowly time varying signal, when examined over a sufficiently short period ( between 5 and 100 msec), during which its characteristics are fairly stationary. But over long periods of time (of the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being produced.

There are several ways to characterize the speech signal. First, is to use a three state representation in which the states are:

i) Silence, where no speech is produced

ii) Unvoiced, in which the vocal cords are not vibrating,

iii) Voiced, in which the vocals cords are vibrating periodically, resulting in a quassi-periodic waveform.

The segmentation of the waveform into well-defined regions of silence, unvoiced and voiced, however is not exact, as it is difficult to distinguish a weak unvoiced sound from silence, or

a weak voiced sound from unvoiced sounds or even silence. The results in fuzzy boundary locations leading to errors.

During the production of speech our articulatory configuration (vocal tract shape, tongue movement, etc.) often does not undergo dramatic changes more than 10 times per second [11]. Therefore, to model speech, it is required to analyze the short time spectral properties of individual sounds, performed at an interval in the order of 10 milliseconds, and characterize the long time development of sound sequences, in the order of 100 msec, due to articulatory configuration changes [11]. There are many spectral-analysis methods like Discrete(Fast) Fourier Transform (FFT), all-pole minimum phase Linear Prediction (LPC) methods, autoregressive/ moving average models, filter-bank methods. However not all spectral properties are important to the human listener and special auditory models are built to emphasize only those spectral properties that are important for human auditory purposes [12,13]. These are short-time spectral vectors known in speech modeling as observation vectors or simply observations. To view the long time development of sound sequences, the frequencies of the successive spectra are plotted against time to generate what is called a spectrogram. In a spectrogram plot, the spectral magnitude is depicted as dark points on the time frequency axis [11**]**.

Juang, Rabiner and Wilpon in 1987, mentioned in their work [14] that spectral vectors, represented by the so-called cepstrum, is defined as the Fourier transform of the log magnitude spectrum – particularly the log magnitude LPC-model spectrum have several advantages in statistical modeling for speech recognition. Computationally, the cepstrum of a stable all-pole system can be found recursively. For speech recognition, a weighting is generally applied to the LPC cepstrum before further processing [14]**.** A vector, such as the cepstrum that represents a short time speech spectrum is considered an observation of speech.

A discrete-symbol representation of the spectral vector of each frame that results from a classification procedure called spectral labeling, is often used as another type of speech observation. The discrete symbol is obtained by choosing one out of a finite collection of several hundred spectral prototypes. The chosen spectral prototype is the one that is closest (in some well-defined spectral sense) to the input speech spectrum. Statistical modeling is performed on the index sequence of the closest spectral prototypes.

There are many ways to characterize the sequence of sounds-that is, running speech-as represented by a sequence of spectral observations, on a longer time basis. The most direct way is to register the spectral sequence directly without further modeling. If we denote the spectral vector at time t by $O_t$, and the observed spectral sequence corresponding to the sequence of speech sounds lasts from t = 1 to t = T, a direct spectral sequence representation is then simply $\{ O_t \}^{T}_{t=1} = ( O_1 ,O_2,. . . , O_T)$. Alternatively, one can model the sequence of spectra in terms of a Markov chain [11]**.** An explicit probabilistic structure is imposed on the sound sequence representation.

## 4. HMMs

A Markov model is a model which depicts every observable event as a state. Speech signals are normally continuous and it is difficult and sometimes even unnecessary to determine how and when a transition from one abstract speech code to another. Hence an explicit, definitive observation of a state sequence, Q cannot be assumed, i.e. each state cannot be uniquely associated with an observable event. The outcomes or the observations are a probabilistic function of each state. The state sequence is not observable. To make the model more flexible, it is assumed that the outcomes or observations of the model are a probabilistic function of each state. These are known as the Hidden Markov Models. Here the actual state sequence is not directly observable (i.e. hidden), it can only be approximated from the sequence of observations produced by the system. That is why such models are known as *Hidden Markov Models.*

An HMM is a doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols [15] and is specified by:

- *The set of states* $S = \{s_1, s_2, . . . , s_N\}$, and a set of parameters $\lambda = \{\pi_i, A, B\}$
- *The prior probabilities* $\pi_i = P(q_1 = si)$ are the probabilities of si being the first state of a state
- sequence. Collected in a vector $\pi$. (The prior probabilities are sometimes assumed equi-probable, i.e. $\pi_i = 1/N$.)
- *The transition probabilities* are the probabilities to go from state i to state j: $a_{i,j} = P(q_{n+1} = s_j \,|q_n = s_i)$. They are collected in the matrix A.
- *The emission probabilities* characterize the likelihood of a certain observation O, if the model is in state $s_i$. Depending on the kind of observation o we have:
- for *discrete observations*, $x_n \in \{v_1, . . . ,v_K\}$; $b_{i,k} = P(x_n = v_k|q_n = s_i)$, the probabilities to observe $v_k$ if the current state is $q_n = s_i$. The numbers $b_{i,k}$ can be collected in a matrix B.
- for *continuous valued observations*, a set of functions describing the probability densities (probability density functions, pdfs) over the observation space. Emission pdfs are often parametrized, e.g, by mixtures of Gaussians.

The operation of a HMM is characterized by

- The (hidden) state sequence $Q = \{q_1, q_2, . . . , q_N\}$, $q_n \in S$.
- The observation sequence $O = \{O_1, O_2, . . . , O_N\}$.

A HMM allowing for transitions from any emitting state to any other emitting state is called an *ergodic HMM*. The other extreme, a HMM where the transitions only go from one state to itself or to a unique follower is called a *left-right HMM*, as shown in fig 1.

### 4.1 Types of HMM

The HMM model where the transitions can be made from any state in some way to any other state is called an ergodic model [16]. Any state will be revisited with probability one. Models which impose a temporal order to the HMM is called the left-right model, where the state sequence which produced the observation sequence must always proceed from left-most state to the right-most state. Lower numbered states account for observations occurring prior to those for higher numbered states.

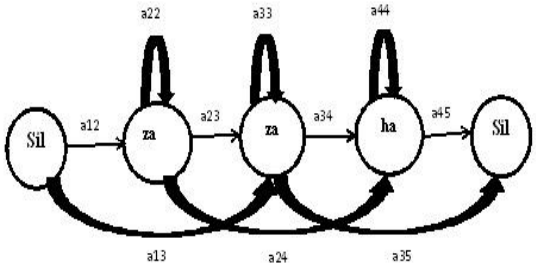


**Fig. 1 A five-state left-right model.**

## 4.2 HMMs for speech recognition

In automatic speech recognition, the task is to find the most likely sequence of words $\hat{W}$ given some acoustic input or $\hat{W}$= arg max P(W|X) for all W an element of *W*. Here, X = $\{x_1, x_2, \ldots, x_N\}$ is the sequence of "acoustic vectors" – or "feature vectors" – that are "extracted" from the speech signal, and we want to find $\hat{W}$ as the sequence of words W (out of all possible word sequences *W*), that maximizes P(W|X). The acoustic feature vectors are the observations and the word sequence are the hidden state sequence of a HMM for speech production.

Words are made of ordered sequences of phonemes: /h/ is followed by /e/ and then by /l/ and /O/ in the word "hello". This structure can be adequately modeled by a *left-right HMM*, where each state corresponds to a phone. Each phoneme can be considered to produce typical feature values according to a particular probability density (possibly Gaussian) (Note, that the observed feature values are d-dimensional vectors and continuous valued).

In "real world" speech recognition, the phonemes themselves are often modeled as left-right HMMs (e.g., to model separately the transition part at the begin of the phoneme, then the stationary part, and finally the transition at the end). Words are then represented by large HMMs made of concatenations of smaller phonetic HMMs [17].

## 4.3 Implementation Issues for HMMs

We find several practical implementation issues discussed in Rabiner's work "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" , 1989. The issues include observation sequences, intitial parameter estimates, missing data, choice of model size and type [16]**.**

- Multiple Observation Sequences. In the left-right or Bakis model form of HMM, the state proceeds for state61 at t=1 to state N at t=T in a sequential manner. The main problem with the left-right models is that one cannot use a single observation sequence for reestimation of the model parameters. Until a transition is made to a successor state, the transient nature of the states within the model only allow a small number of observations for any state. Thus in order to have sufficient data to make reliable estimates of all model parametrs, one has to use multiple sequences.
- Intial estimates of HMM Parameters.The a choice of the intial estimates of the HMM parameters is an important issue, so that the local maximum is the global maximum of the likelihood function. There is no straightforward way to determine this. Experience has shown that either random or uniform initial estimates of $\pi$ and A parameters is adequate for giving useful reestimates of these parameters in almost all cases, and for the B parameters good initial estimates are helpful in the discrete symbol case and are essential in continuous distribution case [16]. The initial estimates of the HMM parameters can be obtained by various ways viz., manual segmentation of the observation sequence(s) into states with averaging of observations within states; maximum likelihood segmentation of observations with averaging; segmental k-means segmentation with clustering, to name a few.
- Choice of Model. Another important issue in implementing HMMs is the choice of the type of model – ergodic or left-right or some other form, the choice of model size indicating the numbers of states, and the choice of observation symbols – discrete or continuous, single of multi-mixture. There is no simple , theoretically correct way of making such choices. The choices are made depending on the signal being modeled.

# 5. SPEECH RECOGNIZERS USING HMMS

Speech recognition systems can be classified into several categories by describing the types of utterances which the algorithm recognizes.

*Isolated Word* recognizers or more appropriately Isolated Utterances, requires a brief pause between a single utterance at a time. An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterance can be a single word or an utterance of words. These systems are said to have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usual processing during the pauses).

*Connected Word* systems or more appropriately 'connected utterances', allow separate utterances to 'run-together' with minimal pause between them. These systems are similar to isolated word recognizers. In the 21word-level stage, each stored word-pattern is matched against all possible regions in the connected word-input pattern. An adjustment window is used to define a region in which each word in the connected word-pattern may start and end for connected speech recognition.

*Continuous Speech* recognizers must be capable of recognizing continuous speech. Here, special methods is required to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally.

*Spontaneous Speech* is natural sounding and unrehearsed. An ASR System with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, some disfluencies like "uhmms" and "aahs", and even slight stutters.

The strengths of the HMM method is its mathematical framework which provides straightforward solution to related problems and its implementational structure which provides flexibility in dealing with various speech recognition tasks and the ease of implementation.

# 6. LIMITATIONS OF HMMS IN SPEECH RECOGNITION

There are also some inherent limitations of this statistical model for speech. Some of the major drawback are reviewed here:

- *The assumption that successive observations are independent.* The successive observations in reality are rarely independent of each other.
- *The Markov assumption itself.* Hidden Markov Modeling is based on the Markov property, which states that the probability of being in a given state at time t only depends on the state at time t-1. This is not always the case for speech sounds where dependencies sometimes extend through several states.
- *The distribution of individual observation parameters can be well represented as a mixture of Gaussian or auto regressive densities.*
- *Constant length observation frames***.** This requirement restricts the possibilities on feature extraction (front end processing). If the frame length be dynamically decided by the front end, better representations could potentially be extracted.
- *Trial and error method for choosing a model topology.* Findings of various researchers show that the left-to-right architecture performs better that ergodic. But there is no formal method for deciding upon the architecture for solving a problem. Also, there is no method to find out the number of states and transitions required for a model, whether to have alternative paths through a model, whether to use the same topology for all the HMM models in that set.
- *The number of parameters needed to set up an HMM is huge.* For a simple four-state HMM with five continuous channels, there would be a total of 50 parameters that would need to be evaluated. 40 of the parameters are means and standard deviations, which are themselves aggregate values.
- *Amount of data required to train an HMM is very large.* As a result of the number of parameters to be estimated in a typical set of HMMs, large training data is hard to be obtained. Sometimes, techniques such as semi-continuous HMMs, triphone clustering and interpolation have been successfully used to improve the adverse effects of insufficient training [18].

Inspite of these limitations they have been found to work well when applied to certain types of speech recognition problems.

# 7. CHALLENGES

- Inconsistencies in the different types of audio and their quality is an important challenge in speech processing. Channel mismatch is one such problem which is the focus of most research in this area. Audio has been gathered using one apparatus and the test audio has been produced by a different channel. The mismatch may be in the handset or recording apparatus; the network capacity and quality; noise conditions; speaker related conditions like illness, stress; transition between media, to name a few [19].
- It is difficult to handle overlapping speech, for example multiple speakers on a single microphone where each speaker is producing similar level of audio at the same time, or in a conference setting, in a room when multiple speakers speak in such a setting. Such captured audio is very hard to recognize.
- Handling of whispered speech, which is very hard to collect as it is not available under natural speech scenarios.

# 8. CONCLUSIONS

HMM is a popular statistical tool for modeling times series data. The variations in speech are modeled statistically using HMM. In speech recognition HMM has been applied with great success to problem as a part of speech classification [20]. This article provides a review of HMM, and its use in speech processing as a powerful tool. Some of the implementation, issues and limitations of speech processing using HMMs were also discussed.

# 9. REFERENCES

[1] Creative Commons Attribution. (2016, February 18). Focal Folds [Online]. Available: http://en.wikipedia.org/wiki/Vocal_folds.

[2] "Human Speech Production Mechansims" , Masaki Honda, NTT Technical Review, Vol.1, No.2, May 2003.

[3] Fant, G., "Glottal source and excitation analysis", Speech Trans. Lab. - Q. Prog. Status Rep. 4, 85-107, 1976.

[4] Titze, I.R., "Physiological and acoustic differences between male and female voices", journal of the Acoustical Society of America, 85, 1699- 1707, 1989.

[5] Stoicheff, M., "Speaking fundamental frequency characteristics of non-smoking female adults", Journal of Speech Hear. Res., 24, 437-441, 1981.

[6] Linke, C.E., "A Study of Pitch Characteristics of female voices and their relationship to vocal effectiveness", Folia Phoniatr, 25,173-185, 1973.

[7] Hollien, H. & Shipp, T., "Speaking fundamental frequency and chronologic age in males", Journal of Speech Hear. Res., 15, 155-159, 1972.

[8] James M. Hillenbrand, M.J. Clark, " The role of $f_0$ and formant frequencies in distinguishing the voices of men and women", 71 (5), 1150-1166, The Psychonomic Society,Inc., 2009.

[9] Ke Wu & Childers, " Gender Recognition from Speech. Part I: Course Analysis", Journal of Acoustical Society of America, 90 (4), 1828-1840, October, 1991.

[10] Peterson,G.E., and Barney, H.L., "Control methods used in a study of the vowels", Journal of Acoustical Society of America, 35, 354-358, 1963.

[11] B.H. Juang and L.R. Rabiner, "Hidden Markov Models for Speech Recognition", Technometrics, Aug 1991, Vol 33, No.3.

[12] Cohen,J., "Application of an Adaptive Auditory Model to Speech Recognition," unpublished paper presented at the 110th meeting of Acoustical Society of America, Nashville, Tennessee, Nov. 4-8, 1985.

[13] Ghitza, O., "Auditory Nerve Representation as a Front-end for Speech Recognition in a Noisy Environment," Computer Speech and Language, 1, 109.

[14] Juang, Rabiner, and Wilpon, " On the use of BAndpass Liftering in Speech Recognition," IEEE transactions on Acoustics, Speech and Signal Processing, 35, 947-954.

[15] Rabiner, Juang, "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, January 1986.

[16] Rabiner, L.R., " A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. of the IEEE, Vol.77, No.2, February 1989.

[17] Barbara Resch,HiddenMarkov Models, "A Tutorial for the course Comutational Intelligence, Signal Processing and Speech Communication Laboratory, Inffeldgasse 16c.

[18] Master's Thesis of Aarnio, Tomi, " Speech Recognition with Hidden Markov Models in Visual Communication", UNIVERSITY OF TURKU, Computer Science, April 1999.

[19] Homayoon Beigi, "Speaker Recognition: Advancements and Challenges", Chapter 1, INTECH, 2012.

[20] Shigeru Katagiri et.al., "A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization," IEEE Transactions on Audio Speech and Language processing Vol.1,No.4.