# Two Pass Hidden Markov Model for Speech Recognition Systems

Waleed H. Abdulla and Prof. Nikola K. Kasabov
Information Science Department
University of Otago
New Zealand
[mailto: waleed@infoscience.otago.ac.nz]

## 1 Abstract

This paper is an approach to increase the effectiveness of Hidden Markov Models (HMM) in the speech recognition field. The goal is to build a large vocabulary isolated words speech recogniser. The model, that we are dealing with, is of continuous HMM type (CHMM). The topology selected is the left-right one as it is quite successful in speech recognition due to its consistency with the natural way of articulating the spoken words.

The main task here is to extract the spoken words from their background using one CHMM and process them in isolation by another models. This is considered as a perceptual way of extracting the signal. This technique is substantially increasing the performance of the system and improving the incorporation of states' duration.

## 2 Introduction

The one of many objectives in implementing speech recognition systems is to separate the signal of essence from the background environment as faithfully as possible. This operation has crucial effect on the overall performance of the recogniser. It is an issue to be tackled by the researchers from the early beginning of this field. The early milestone technique was using explicit features for speech non speech discrimination; such as speech signal energy and zero-crossings[1,2,7]. This technique is effective in case of low noise environment, but unreliable with the increasing noise and varied articulation manners such as breathing and clicks. The other approach was the pattern classification of voiced, unvoiced, and silence states[3, 4]. This technique implies some decision making to improve the performance of the system but it incurs a heavy computational load. Hybrid techniques were also suggested to alleviate the computational load while improving the performance. Wilpon et al. benchmarked a multispeaker digit recogniser to evaluate the effect of misaligned word boundaries on the recognition rate. The words and the reference patterns were manually extracted. The recognition rate was found to be 93%; which was the utmost value. Then misalignment procedure was practised with recognition error measured at each step. Fig.(1) shows the contour plot of the performance variation over the end points perturbation[4]. Recent techniques dealt with presilence and postsilence periods as pre and post states of Hidden Markov Models (HMM). During training phase the words are modelled without including the silence periods, while the silence periods are modelled as separate states. In recognition phase the pre and post silence states are concatenated to the initial and final states of the words' models. Then the maximum likelihood ( or any other optimisation ) procedure are followed to identify the tested words. Those HMM techniques, even they are effective, still need to concatenate the silence states during recognition phase that consequently increases the computational cost, especially with long silence periods and increasing number of models. Other successful techniques are using neural networks (NN) to model the silence periods, but these need decision making steps to identify the positioning and the relevance of the detected silence periods[5].

This paper looks at the problem from different point of view which makes use of the early ideas of deleting the silence periods and recent ideas of modelling them with HMM. A method is suggested that shows superiority over the other techniques and translates it into great potential in recognition performance.

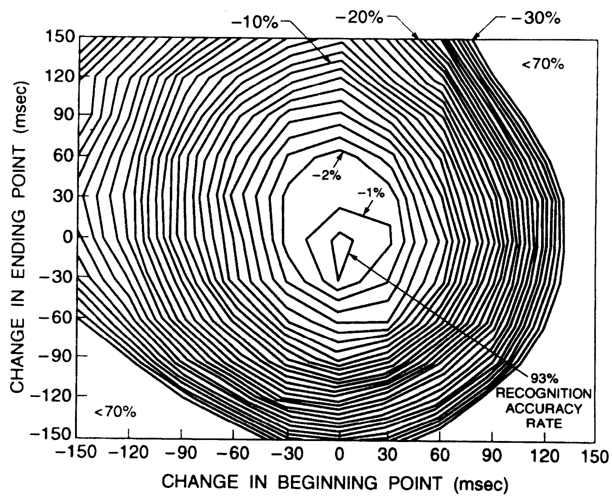


Fig.(1) Recognition performance contour showing the sensitivity of the digit recognition system to the words end points variations.(after Wilpon et al.[4].

## 3 System Modelling

The main modelling here is based on Continuous Hidden Markov Modelling (CHMM) technique[6,7,8,9]. During training a model is constructed from a collection of words, including their silence periods. The model learns only the first and last states of words as they represent the pre and post silence periods. Throughout the recognition phase the candidate words are first aligned with this all words model

and the speech samples belonging to the first and last states are removed. Then the extracted speech segment is aligned with the different words' models to select the most probable model to issue the spoken word.
The modelling of the system implies two steps: training and recognition.

## 3.1 Training

There are two types of models to train - word extraction model, and word recognition models.

### 3.1.1 Word Extraction Model

The data set here consists of words from many different words spoken in different background environments (not necessarily all the words of the recogniser), and normally selected in the range of 50-100 words.
The model parameters are trained for left-right CHMM with 7 states and one mixture (unimodal). Multimixtures could be used here but it does not show more difference in performance over the unimodal one. Fig.(2) shows the structure of the model used, it shows just five states for demonstration. The observations are Mel scale coefficients of the speech signal frames with only 13 coefficients (12 mels plus one power coefficient). The delta coefficients are not included to make the model insensitive to the dynamic behaviour of the signal and then gives more stable background detection. The speech frames for building the model are selected to be 11.6 ms taken each 2.9 ms. This model is called an all-words model due to its way of training.

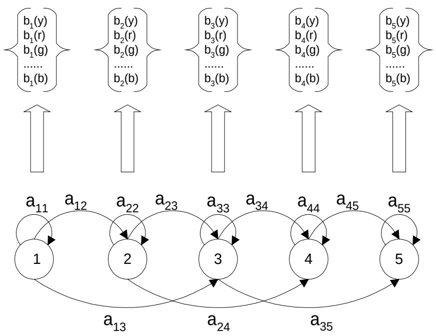


Fig.(2) CHMM left-right topology. Where $a_{ij}$ is the state transition probability from state i to j, $b_k(o)$ is the observation probability function in state k.

### 3.1.2 Word Recognition Models

This is the process of building a model for each spoken word. In this stage 50-100 utterances of the same word are taken from different speakers to perform the data set . The training data are taken from Otago Speech Corpus, which is freely available from the Internet on

http://Kel.otago.ac.nz/hyspeech /corpus

The observation sequence in this case are the mel scales with 39 coefficients (12 mels and one power with their deltas and delta-delta coefficients). This makes the model more sensitive to the dynamic behaviour of the signal which in this step is the main objective of modelling. Regarding the other parameters, the topology and the type is left-right CHMM as in Fig.(2), with 9 states and 12 mixtures. The speech frames in this case are of 23 ms taken each 9ms.
The probability density function (pdf) of certain observations $O$ being in a state is considered to be of Gaussian Distribution .
The general form of $b_i(O)$ is:

$$b_i(O) = \sum_{m=1}^{M} c_{im} \aleph(O; \mu_{im}, U_{im}), \quad 1 \le i \le N$$

where:

$c_{jm}$ : is the m-th mixture gain coefficient in state $i$.
$\aleph$ : is the pdf distribution which is considered to be Gaussian in our case.
$\mu_{im}$ : is the mean of the m-th mixture in state $i$.
$U_{im}$ : is the covariance of the m-th mixture in state $i$.
$O$ : is the observations of feature dimension d.
$M$ : is the number of mixtures used.
$N$ : is the number of states.

In the optimisation procedure to find the best mixture distribution during training step, the Vector Quantization (VQ) technique is applied on the unimodal data set of each state. The observations belonging to each state are extracted by Viterbi Algorithm during training and then optimised by using the maximum likelihood method. This representation results in a good modelling of the data. The following section explains this technique in more details. During this training step a model for each word is built in addition to one model for word extraction.
The state duration factor is incorporated through using heuristic technique, which boosts the performance to the same level as the correct theoretical duration inclusion with very low computational and storage costs. The state duration probability function $p_j(\tau)$ is estimated during the model training and defined as:
$p_j(\tau)$ : is the probability of being in state j for $\tau$ duration, normalised to the length of observations.
The duration probability density function is considered to be Gaussian with 5 mixtures.

### 3.1.3 Mixture Density Components Estimation using Maximum Likelihood (ML):

The ML estimation is an optimisation technique that can be used efficiently in estimating the different component of multimixture models. We are not going through the mathematical derivations of the ML but only describe the method to be used in our task.
The following definitions are used further in the paper:
$b_i(O_t)$ : probability of being in state i given observation sequence $O_t$. It is considered to be Gaussian distribution .
$c_{im}$ : probability of being in state i with mixture m (gain coefficient).
$b_{im}(O_t)$ : probability of being in state i with mixture m and given $O_t$.
$\Phi(w_{im}|O_t)$ : probability function of being in a mixture class $w_{im}$ given $O_t$ in state i.
$T_i$ : total number of observations in state i .
$T_{im}$ : number of observations in state i with mixture m.
N : number of states.
M : number of mixtures in each state.

Now we are ready to implement the algorithm through applying the following steps:

1 – Take several versions of observations of certain word, say digit "zero", spoken several times by many speakers.

2 – Apply standard CHMM using unimodal representation ; then via Veterbi algorithm detect the states of each version of the training spoken word.

3 – Put the whole observations belonging to each state from all the versions of the spoken word into separate cells. Now we have N cells and each one represents the population of certain state derived from several observation sequences of the same word.

4 – Apply vector quantization technique to split the population of each cell into M mixtures and getting $w_M$ classes within each state.

5 – Use any of the well known statistical methods to find the mean $\mu_{im}$ and the covariance $U_{im}$ of each class. The gain factor $c_{im}$ can be calculated by:

$$c_{im} = \frac{\text{number of observations being in state i and mixture m}}{\text{total number of observations in state i}}$$

6 – Calculate $\Phi(w_{im}|O_t)$ from the following formula:

$$\Phi(w_{im} \mid O_t) = c_{im} . \frac{b_{im}(O_t)}{b_i(O_t)}$$

7 – Find the next estimate of $\hat{c}_{im}, \hat{\mu}_{im}$, and $\hat{U}_{im}$ from the formulas given by ML :

$$\hat{c}_{im} = \frac{1}{T_i} \sum_{i=1}^{T_i} \Phi(w_{im} \mid O_t)$$

$$\hat{\mu}_{im} = \frac{1}{T_{im}} \sum_{i=1}^{T_i} \Phi(w_{im} \mid O_t).O_t$$

$$\hat{U}_{im} = \frac{1}{T_{im}} \sum_{i=1}^{T_i} \Phi(w_{im} \mid O_t).(O_t - \hat{\mu}_{im})(O_t - \hat{\mu}_{im})'$$

$$\hat{b}_{im}(O_t) = \sum_{m=1}^{M} \hat{c}_{im} \aleph(O; \hat{\mu}_{im}, \hat{U}_{im}), \quad 1 \le i \le N$$

$$\hat{b}_i(O_t) = \sum_{m=1}^{M} \hat{c}_{im} \hat{b}_{im}(O_t)$$

8 – Compute the next estimate of $\Phi$ by using the formula:

$$\hat{\Phi}(w_{im} \mid O_t) = \frac{\hat{c}_{im}\, \hat{b}_{im}(O_t)}{\sum_{n=1}^{M} \hat{c}_{in}\, \hat{b}_{in}(O_t)}$$

9 – IF $|\Phi(w_{im} \mid O_t) - \hat{\Phi}(w_{im} \mid O_t)| \le \varepsilon$ THEN END

ELSE Make the new value of $\Phi(w_{im}|O_t)$ equal the newly predicted one.

$$\Phi(w_{im} \mid O_t) = \hat{\Phi}(w_{im} \mid O_t)$$

GO TO STEP 7.

Here $\varepsilon$ is a very small threshold value.

## 3.2 Recognition

This step comprises two operations:

1- The input unknown utterance is submitted to the first model - word extraction model, to extract efficiently the spoken word from the background.

2- The extracted word from the previous operation is submitted to all the other models. The model that scores maximum log likelihood $\log[P(O/\lambda)]$ is representing the submitted input, where $P(O/\lambda)$ is the probability of observation O given a model $\lambda$.

The duration factor is incorporated through an efficient formula which results in improved performance.

During recognition, the states' duration are calculated from the backtracking procedure in Viterbi Algorithm. Then, the log likelihood value is incremented by the log of the duration probability value as shown below:

$$\log[\hat{P}(q, O \mid \lambda)] = \log[P(q, O \mid \lambda)] + \eta\, T \sum_{j=1}^{N} \log[p_j(\tau_j)]$$

where: $\eta$ is a scaling factor;
$T$ is the length of the observation sequence;
$\tau_j$ is the normalised duration of being in state j as detected by Viterbi Algorithm.

## 4 Results

To build a robust word recognition model different effects have been included in the silence periods. The training pre- and post silence periods include: mic clicks, sound artefacts, lip slaps, heavy breathing. The collected all words should include, in their start and end phones, the most problematic phones such as weak fricatives, weak plosives and nasals. The best performance is achieved by inclusion as many effects as possible from different speakers.

The system was benchmarked against several other techniques and shows tractable results in determining the actual start and end points (boundaries) of the tested words. The following figures show clearly how the system works in extracting the spoken word of digit "five" from the background. This example is chosen as it suffers from bad extraction using the other techniques due to low energy starting fricative state as well as fading end. The background noise level is taken to be comparable to the start and end levels of the tested word to make it as difficult to detect as it could be. Fig(3) shows the boundary of the signal using an explicit technique. The result is troublesome for the recogniser. Fig.(4) shows the same spoken utterance as detected by all-words model. The extracted signal (after removing the samples belong to states 1 and 7) is submitted to the available models of the words for recognition. The time signal and spectrogram are displayed to show the correspondence between the signal and the states, which indicate how precise the model is. The states are detected through the backtracking phase in Viterbi Algorithm. One more experiment which consolidates the technique is by looking at the log likelihood of the observation sequence as presented to a single state silence model. Fig(5) shows the silence states in the spoken word six. The pre, mid and post silence periods are clearly noticed from the highest probability levels.

The recognition rate using the technique described in this paper is scoring more than 95% when tested by twenty four persons, speaking the digits words.

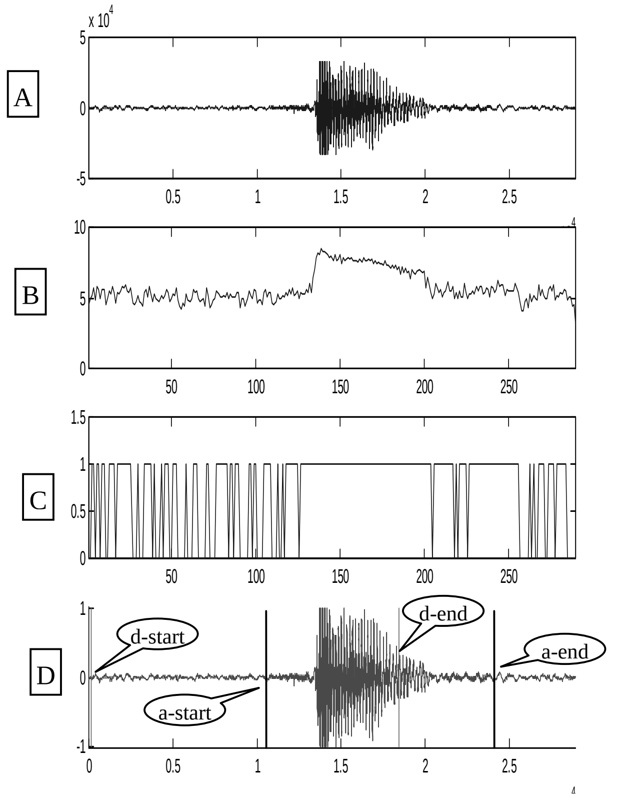


Fig.(3) End points detection using explicit techniques
A: The input signal, B: The energy level of the signal, C: Over-Threshold energy level, D: Normalised signal with word boundary detected.
d-start and d-end for detected bounds, a-start and a-end for actual bounds.

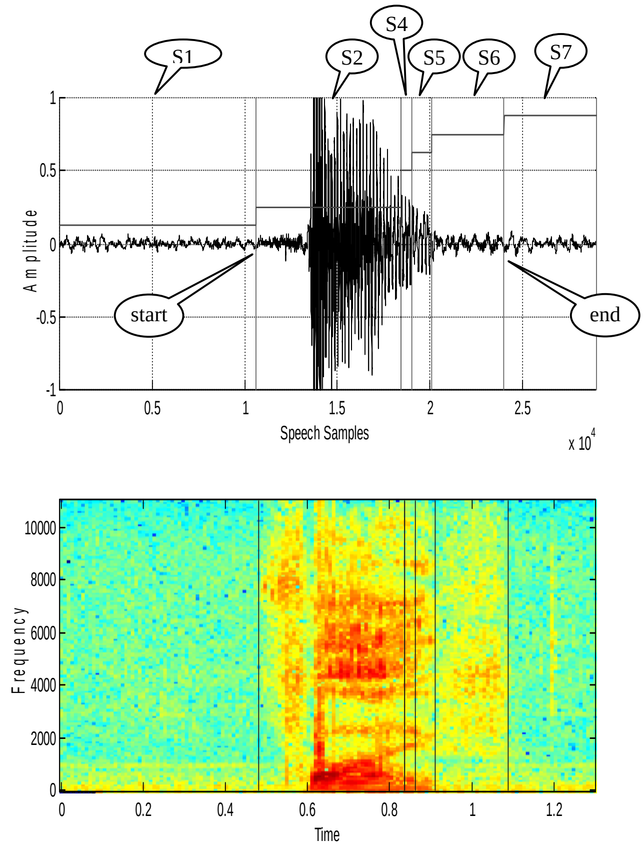


Fig.(4) End Points Detection using all-words Model.
The start and end points detected match the actual ones. The spectrogram shows the spectra of pre and post silence as well as the speech signal. The relevant states $S_i$ are the first and the last ones. Some states might be jumped as in state3 (S3) in this example.

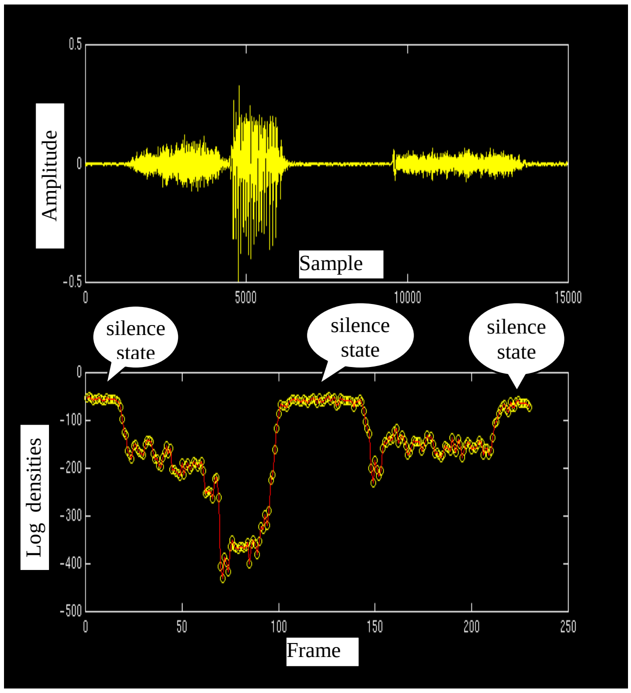


Fig.(5) Silence states in the spoken digit "six".
The silence states are identified by their high log probability densities.

## 5 Conclusions

The technique presented in this paper offers an efficient way of extracting the speech signals from their backgrounds. It shows superiority over the known techniques of end points detection as it is a perceptual way which takes into consideration the input signal as well as the background status in taking the decision. This does not incur further or more computational cost as it might appear from the first look. The word extraction model will save at least 1/3 of the computations as the extracted signal has shorter duration than the original one (signal plus silence periods).
The CHMM could be applied on the input signal without performing word extraction but the computation in this case will be more as the duration of the signal will be longer. Accordingly, the number of states will be more to compensate for the background states. The known end points detection methods degrade the performance of the system specially in the case of low energy segments at the beginning and/or at the end of the speech signal. The all_words model which is used to extract the words from background environment is flexible and could be easily adapted to any environment just by presenting the new environment during training.
The precise signal /background separation leads to high recognition rate. The post inclusion of the normalised states' duration in the log probability equation using the way described in section 3.2 adds further reinforcement to the performance of the system to raise the recognition rate to more than 95% with speaker independent digits data set.

## 6 References


[1] Lawrence R. Rabiner and M. R. Sambur. " An Algorithm for Determining the End Points of Isolated Utterances", Bell System Technical Journal (BSTJ), no. 54, pp 297-315, Feb 1975.

[2] Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg, and Jay G. Wilpon. " An Improved End Points Detector for Isolated word Recognition", IEEE ASSP , vol. 29, no. 4, pp 777-785, Aug 1981.

[3] Bishnu S. Atal and Lawrence R. Rabiner. " A pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Application to Speech Recognition", IEEE ASSP vol. 24, no. 4, pp 201-212, June 1976.

[4] J. G. Wilpon, L. R. Rabiner, and T. B. Martin, "An Improved Word –Detection Algorithm for Telephone Quality Speech Incorporating both Syntactic and Semantic Constraints, " AT & T Tech. J. vol. 63, no. 3, pp 479-498, March 1984.

[5] N. K. Kasabov, "Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering", Cambridge, MIT Press, 1996, Chap. 5.

[6] L. R. Rabiner and Biing-Hwang Juang, " Fundamentals of Speech Recognition", Printice Hall Signal Processing Series, Alan V. Oppenheim , Series Editor, 1993, Chap. 4.

[7] J. R. Deller, J. G. Proakis, and J. H. Hansen, "Discrete-Time Processing of Speech Signals", New York: Macmillan Publishing, 1993, Chap. 4, 12.

[8] W. H. Abdulla and N. K. Kasabov, "The Concepts of Hidden Markov Model in Speech Recognition", Technical Report TR99/09, University of Otago, July 1999.

[9] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, Feb. 1989.