A GEOMETRIC INTERPRETATION OF HIDDEN MARKOV MODEL

Chee Wee Loke, Roberto Togneri
Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
The University of Western Australia

Abstract -In this paper, we investigate the relationship between speech trajectories and the hidden Markov model. The speech utterances were transformed into speech feature vectors and the trajectories displayed on a two dimensional space. The hidden Markov models were also displayed on a two dimensional space. By visual examination, we think that the state of the HMM is related to a sustained sound. Further experiments showed that each state seem to be associated with a distinct phoneme of the utterance. Therefore, the number of states required in the continuous HMM is related to the number of phonemes in the word to be modelled. In the semi-continuous HMM, it is also possible that the same gaussian probability density function is shared by the same phoneme sound in different semi-continuous HMMs.

1.0 INTRODUCTION

Recently, hidden Markov modelling (HMM) (Rabiner, 1989) has been widely used in many state-of-the-art speech recognition systems. Hidden Markov modelling is a doubly stochastic process with an underlying stochastic process that is not observable, therefore, it is hidden. To illustrate this point, suppose we build two hidden Markov models to model the words "FIVE" and "NINE". At the end of the process, we can only see the outcome after comparing the testing vectors with both models sequentially. If the testing vectors corresponding to the word "FIVE", then HMM for the word "FIVE" will return the highest probability. However, what is exactly happening in each state is not known or it is hidden.

In this study, we used a visual tool called *fview* (Lee & Alder, 1993) developed by the Centre for Intelligent Information Processing Systems to display features produced from a specific front-end in an interactive fashion. We employed *fview* to explore the relationship between speech feature vectors and the corresponding states of the HMM. This is done by displaying speech feature vectors together with the states of that hidden Markov model.

This is shown in Figure1. A 3-state HMM is displayed together with the trajectory of an utterance. The trajectory represents finite dimension feature vectors of speech utterance "FIVE". By visually examining the trajectory and states of the HMM, we think that each state actually corresponding to a sound, maybe a phoneme. We investigated two different hidden Markov models, the continuous hidden Markov model (CHMM) and semi-continuous hidden Markov model (SCHMM). The results indicated that each state corresponds to a particular sound.

2.0 PREPROCESSING AND PROJECTION

All feature vectors were generated as FFT filterbank coefficients. A FFT over 512 points with a 200 point advance was applied to the speech data, previously digitized at a frequency of 16 MHz. For every 512 point frame the resulting values were reduced to 12 filterbank coefficients using a simulated filterbank of 12 overlapping filters. Each frame is represented by fixed number of values. In order to project 12 dimensional vectors to a 2 dimensional space, we need to define a plane in that space. Projecting all the data points

onto this plane will give a two dimensional view through the space. Continuous line connecting the data points is referred to as trajectory.

The second object to be displayed is the hidden Markov model. The hidden Markov model is a finite-state model. Each state is characterized by a finite-dimension mean vector and covariance matrix. A finite dimension gaussian probability density (PDF) function can be projected down to 2 dimensional space and displayed on two dimensional space. In the continuous HMM, the model is characterized by a finite number of states. In Figure 1, the clusters labelled as State1, State 2 and State 3 are the states of the HMM.

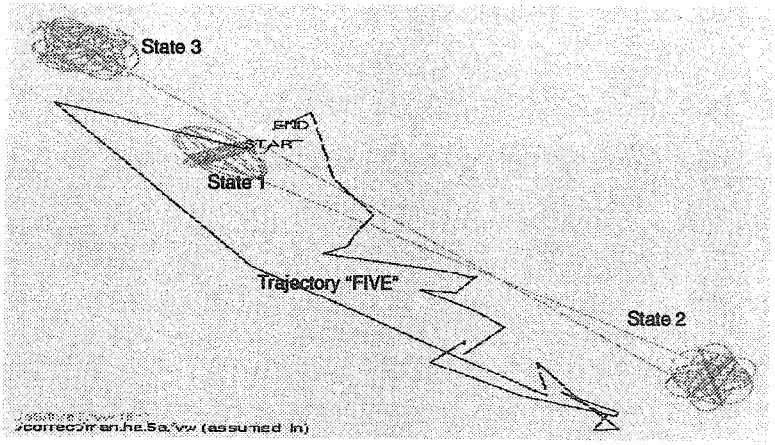

Figure 1. Example of HMM states and trajectory for the word "FIVE".

3.0 EXPERIMENTAL EVALUATION

3.1 Continuous Hidden Markov Model

In the continuous hidden Markov model (CHMM), parameter estimation is based on the maximum-likelihood methods on the assumption that the input data have indeed been generated by a gaussian process. CHMM consists of a number of emitting states, each state is defined by one or more gaussian mixture components each with a weight; and finally each mixture component is defined by a mean vector and a covariance matrix. In this paper, each state is characterized by one gaussian mixture component.

The issue of the number of states to be used in each word model leads to two different ideas. One is to specify the number of states corresponding roughly to the number of sounds (phonemes) within the word. The second idea is that the number of states should correspond roughly to the average number of feature vectors in a spoken version of the word, this is called the Bakis model (Bakis, 1976). Our experimental results show that the idea of states correspond to phonemes is a more accurate word modelling technique.

In these experiments, 10 HMMs were built using 2-state, 3-state, 5-state and 7-state models. These are used to model ten digits. The results are shown in Table 1. It shows that by increasing the size of the HMM from 3 to 5 or 7 does not increase the accuracy very much. We can conclude that to model the word

"ZERO" to "NINE", a 3-state model is sufficient.

TABLE 1. Recognition Accuracy with varying HMM size

Number of States	2-state	3-state	5-state	7-state
Accuracy (%)	91.26	97.89	98.49	98.59

Figure2 shows the trajectory with its feature vectors labelled by the frame sequence numbers together with the 3-state CHMM for the word "FIVE".

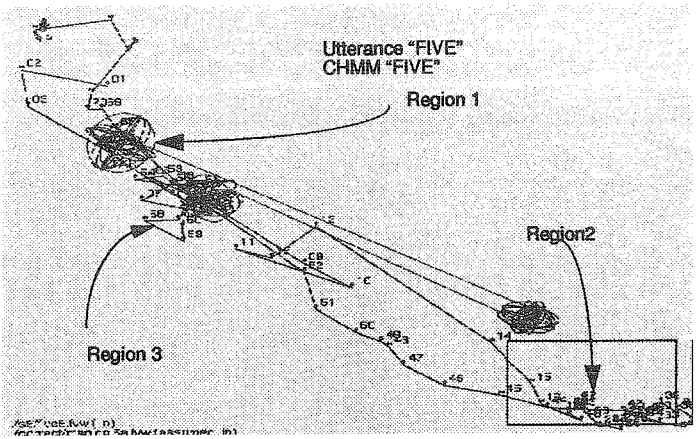

Figure 2. Clusters and States

The trajectory stays around region 1 for 12 frames and then quickly moves or "jumps" into region 2 at the bottom right hand corner. It seems to stay in region 2 for a long period of time, then the trajectory moves into the last region, which is labelled as region 3. We can see that there are 3 clusters. The location of the clusters correspond to the states of that particular HMM in the feature space. The next step is to find out what is actually embedded in each state. One possible hypothesi is that each state representing a particular sound, maybe a phoneme. In our experiments, we found that each state is likely be related to a particular phoneme.

In Figure 3, CHMM for the word "NINE" is displayed. "NINE" consists of 3 phonemes, /n/,/ay/, /n/. If each state corresponds to a phoneme, then we would expect to see that the beginning and end states are close to one another because they are of the same phoneme. The beginning state (State 1) is very near to the end state, labelled as State 3. In Figure 3, the CHMM for the utterance "ONE" is also displayed. Utterance "ONE" consists of the phoneme sequence /w/, /ah/, /n/. The last state of the CHMM of "ONE" (State C) is located very close to State 1 and State 3 of CHMM of "NINE". It could be due to the nasal /n/ sound.

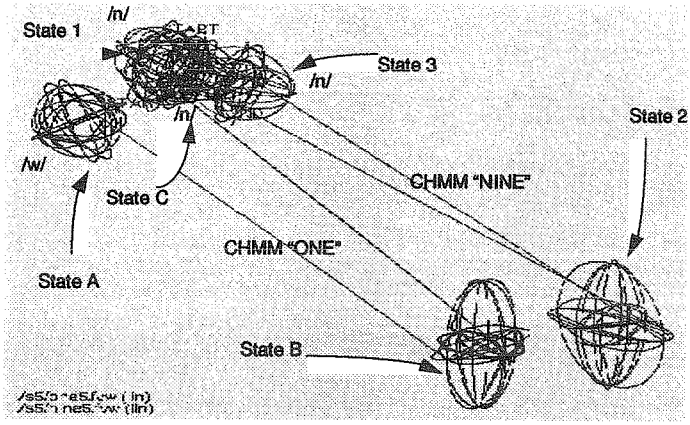

Figure 3 CHMM for "ONE" and "NINE".

Further investigation was carried out to confirm the results. From Viterbi algorithm, information about the probability of feature vector belong to certain state was obtained. An utterances were then manually segmented and played back. It again confirmed the results presented above. Each state of the Markov model corresponding to a particular sub-unit of a word or utterance. Building a particular HMM, we are segmenting each utterance into its sub-unit.

The majority of the utterances are made up of 3 phonemes. It is not surprising that we only need a 3-state model to represent utterances of the ten digits. For utterances with larger number of phonemes, such as "SEVEN" and "ZERO", 5-state CHMMs give almost 3% improvement in terms of recognition accuracy. Experimentally, we have shown that the idea of choosing the number of states corresponding "roughly" to the number of sounds (phonemes) within the word seems more appropriate.

In CHMMS for words "FIVE and "NINE", the states are very close to one another. This is reflected in Table 2, where the recognition accuracy for "FIVE" and "NINE" are the lowest. If we can model the transition from one state to another, say, transition for "FIVE" is different from "NINE", then we can also tell them apart. In this study, we find that there are not many feature vectors between clusters, there is not much information we can extract from the transition stage. We already saw that the trajectory stays in a region for a while and then very quickly "jumps" to a new region. There are very few points in between two states.

TABLE 2. Recognition for digits "ZERO" to "NINE".

Utterance	Accuracy (%)	Utterance	Accuracy (%)
ONE	99.55	SIX	99.11
TWO	98.66	SEVEN	98.66
THREE	99.11	EIGHT	98.21
FOUR	99.55	NINE	97.77
FIVE	97.66	ZERO	98.66

3.2 Semicontinuous Hidden Markov Model

The major disadvantage of the CHMM is the computational complexity. To reduce computations, semicontinuous hidden Markov model (SCHMM) (Huang & Jack, 1989) is used. The SCHMM maintains the modelling ability of large probability density functions (PDFs) by using a universal set of PDFs. The number of free parameters and the computational complexity can be reduced because all of the PDFs are shared across different models. In another words, we put all the covariance matrices into a pool and these are shared by all the HMM models. If what we claimed is true, then, each PDF in the pool should also represent a specific sound (phoneme). We require roughly 24 PDFs to model the 10 digits since there are roughly 24 distinct phonemes in utterances of the ten digits. From Table 3, we notice that by increasing the pool size, significant improvement is achieved in terms of accuracy, but not much after the pool size reaches 24.

TABLE 3. SCHMMs vs. accuracy

Pool Size	12	18	21	24	30	64	128
Accuracy(%)	91.03	94.76	95.45	97.89	97.52	97.85	97.89

Comparing with continuous HMM, which need 33 PDFs, we can save 9 PDFs. To show that each PDF in the SCHMM pool is actually representing a phoneme, we plot the CHMM together with the SCHMM states on the screen.

In Figure 4. the CHMMs for words "FIVE" and "NINE" are displayed. In SCHMM the models for "FIVE" and "NINE" share the same gaussian that is shown by the state labelled as [SCHMM,/ay/]. It is clear that all three gaussians (["NINE",CHMM,/ay/], ["FIVE", CHMM, /ay/] and [SCHMM,/ay/]) are located in the same region of the speech space. This seems to show that they correspond to the same phoneme, the /ay/ sound.

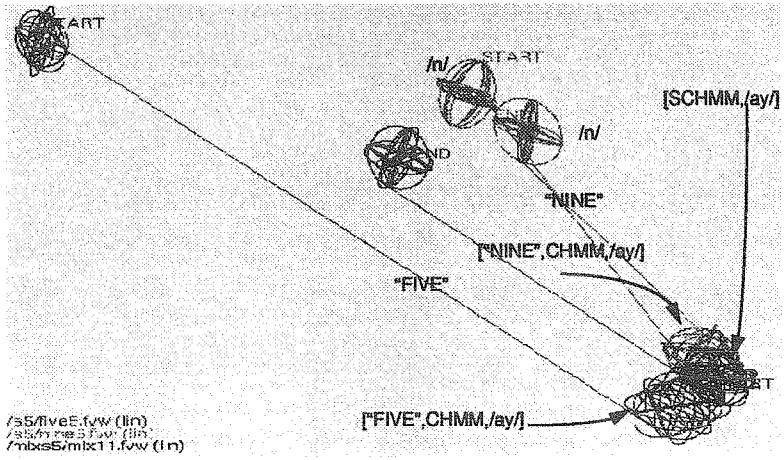

Figure 4. SCHMM and phoneme.

4.0 CONCLUSION

This paper presents a new way of looking at an HMM recognition system. Previously, we can only look at the outcome or the results but do not know what is happening in between. By displaying the states of the HMM together with the corresponding utterances, we can reveal some of the "hidden" parts of hidden Markov model. By visually examining the trajectory and PDFs, it appears that each state actually corresponding to a sound, maybe a phoneme.

5.0 REFERENCES

Bakis,K. (1976), *Continuous speech word recognition via centisecond acoustic states*, in Proc. ASA meeting (Washing,DC).

Huang,X.D. & Jack,M.A. (1989), *Semicontinuous hidden Markov models for speech recognition*, Computer Speech and Language 3 pp. 239-25.

Lee,G. & Alder,M. (1993), *A Geometric Interpretation of Speech Features*, First Australian and New Zealand Conference on Intelligent Information Systems, Perth, Australia.

Rabiner,L.R. (1989), *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proc. of the IEEE, Vol. 77, No 2, February 1989, pp 257-258.