A Visualization Tool for Continuous Density Hidden Markov Models

Sunil Kopparapu
Cognitive Systems Research Laboratory

November 19, 2003

Abstract

Continuous density hidden Markov models (CD-HMMs) are doubly stochastic processes which are extensively used in speech signal processing. In the case of speaker independent isolated spoken word recognition systems, the spoken words are usually modeled using HMMs. Given a large number of speech samples of the word, it has been found that HMM can model the word very well. The *good fit* of HMMs to model speech is strengthened by the fact that there has not been a paradigm shift since after hidden Markov models (HMMs) have been used in the area of speech recognition or speaker verification.

While continuous density HMMs are in extensive use, to most of the speech community the HMMs remain abstract in the sense there has been no *nice* way of visualizing the HMMs. CD-HMMs are typically represented by (i) number of states, (ii) state transition probabilities and (iii) a Gaussian distribution, represented by μ (mean) and σ^2 (variance) for each state. In this paper, we present a methodology to visualize an HMM. These visuals serve two purposes, they give, especially for a novice in the area of speech technology a *feel* for the HMMs and secondly, the HMMs of words can be visually compared quickly to check if the words are modeled *similarly*, because words having similar models can lead to confusion. This in turn effects the speech recognition accuracy. The HMM visualization tool, developed using open domain tool (octave, gnuplot) captures all the HMM parameters.

1 Introduction

HMMs are statistical models and are widely used because they have proved effective in a number of domains and for numerous applications (for example [1, 2, 3]). The most significant of these is the use of HMM in most commercial systems speech recognitio systems. They have also proved effective for a number of other tasks, like speech synthesis, character recognition and DNA sequencing, handwriting recognition, natural language processing [4], molecular biology[5] and sign language recognition [6] to name a few.

In speech signal processing, in general, the given acoustic speech waveform is first segmented into shorter segments called frames. Typically a frame is made up of 20 ms of speech data; this segmentation is necessary because speech signal is non-stationary and it is assumed that 20 ms of speech data is in general stationary and hence much of the signal processing techniques can be applied. The frames are moved over the full speech sample such that the frames overlap; typically the overlap is 10 ms. Let N be the total number of frames in a given speech sample. Each frame (say $k = 1$ to N) is analysed separately and *feature vector*

(f_k) extracted from the speech signal under each frame. Typically the length of the feature vector f_k is the number of features used to represent the speech signal. For example, if 10 MFCC (Mel Frequency Cepstral Coefficients), 8 Δ-MFCC and 4 Δ^2-MFCC are the features used to represent the speech signal within the frame, then f_k, the feature vector of the k^{th} frame has dimension 22. Let the complete feature vector whic represents the speech signal be denoted by $\mathcal{F}$ which is obtained by concatenating the individual feature vectors, namely $\mathcal{F} = [f_1 f_2 \cdots f_N]$. Such $\mathcal{F}$'s obtained from several repetition of the same word (typically 100 or more) forms the input set to train (develop) an HMM for that word, note that each speech sample will have different number of frames.

Training aspects of HMMs is very well cited in literature (for example [7, 8]) and hence we will not dwell into the training aspects of HMMs. We will concentrate on the parameters that capture the HMM representation of the word. HMMs are typically represented by

1. η, the number of states
2. $\mathcal{T}$, the state transition probability matrix of size $\eta \times \eta$ and
3. $\mathcal{G}_k(\mu, \sigma^2)$, for each state $k = 1, \cdots \eta$, a set of η Gaussian distribution, with mean μ and variance σ^2.

If there are N features then there would be N $\mathcal{G}_k(\mu, \sigma^2)$. For example for the feature l and state m, the Gaussian would be represented by $\mathcal{G}_m^l(\mu, \sigma^2)$.

> **Note:** *$\mathcal{G}_m^l(\mu, \sigma^2)$'s for $l = 1, \cdots N$ and $m = 1, \cdots \eta$ together with η, the number of states and $\mathcal{T}$, the state transition probability. represent hidden Markov model of a spoken word completely.*

In the next section we give a visual presentation of the HMM. The basic idea of building a visualization tool is to capture *all* the details of an HMM.

2 Visual Representation

The need for developing a tool for visual representation of an HMM is important for several reasons. The foremost being, HMMs have been used extensively in literature but there is no *feel* for what they look like or how the HMM of one word is different from the HMM of another word. In our opinion, a visualization tool would enable us to 'see' the HMM and also help make visual comparision between HMMs. An HMM visualization tool is very useful, for example, it could be used as

- a tool for a novice in the field of speech recognition, to understand what an HMM is like without having to use them; as it stands today, HMMs are the most powerful tool to model speech.
- for an isolated word speech recognition developer, visual comparision of HMMs would bring out clearly if the spoken words are trained well or there is a scope for confusio,n even before actually testing to find the words that can get confused because their HMM are *similar*.
- visualization of HMM it also shows if the *training* is sufficient or the HMMs required to be further tuned by exposing it to more data.

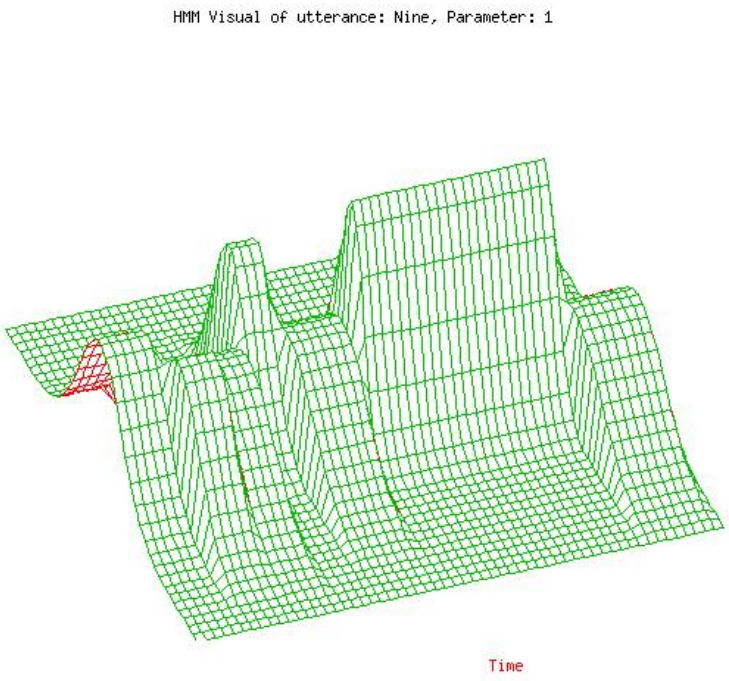


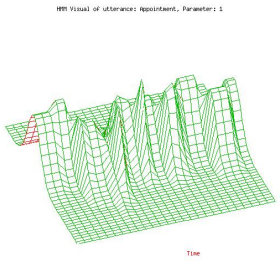
Figure 1: HMM corresponding to the first parameter of the utterance *nine*.

The HMM visual representation of the utterance *nine* is shown in Figure 1. There are 7 visible *gaussian bands* in Figure 1 which correspond to the 7 states of the HMM (the HMM actually has 9 states, but the first and the last states are not considered because they are emitting states). The length of the gaussian band along the time axis captures the state duration, for example it can be seen that the 6^{th} state has the longest duration. This length of the band captures the state transition probabilities. The location of each gaussian band captures the μ of the Gaussian in that particular state (note that μ, the mean, of the Gaussian within a state is constant). The center of the axis perpendicular to the time axis represents $\mu = 0$. Clearly, the second band (state 2) has a negative mean, while the state 4 has a positive μ. The spread of the band captures σ, the variance of the Gaussian. For example the Gaussian corresponding to the state 5 (band 5 in Figure 1) has a small variance (small spread) while the 7^{th} band has a large variance (larger spread). It is clear that Figure 1 is able to capture *all* the details of pertaining to the HMM of the utterance *nine*[1].

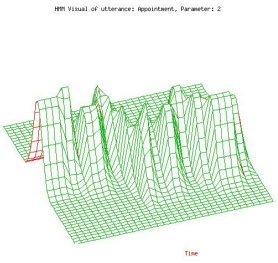
If N parameters are used to represent a speech frame, then we would have N such visual representations corresponding to each parameter for a given word. For example, Figures 2 and 3 show the HMM's for the words *Appointment* and *Sunil Kopparapu* respectively for the first 8-MFCC parameters.

Clearly the HMM visuals (compared parameter wise) of the two spoken words are very distinct and hence one can be sure that there will be no confusion between the spoken word *Appointment* and *Sunil Kopparapu*. This is one of the benefits of being able to visualize HMMs for any word.

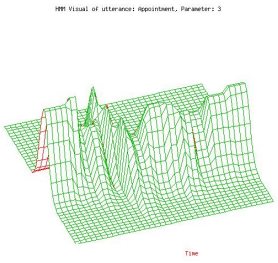
[1]Note that a complete representation of the utterance *nine* would have been a sequence of such 3D plots for all the parameters.



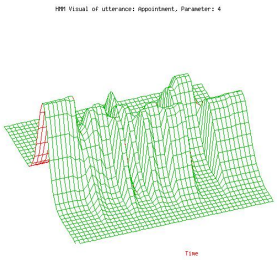
(a) Par 1



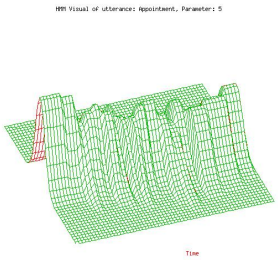
(b) Par 2



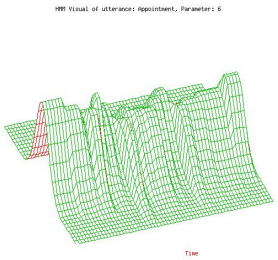
(c) Par 3



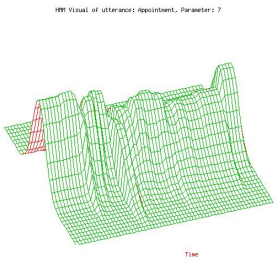
(d) Par 4



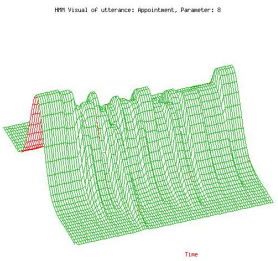
(e) Par 5



(f) Par 6

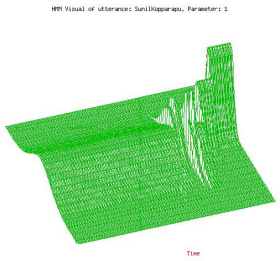


(g) Par 7

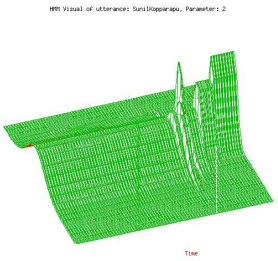


(h) Par 8

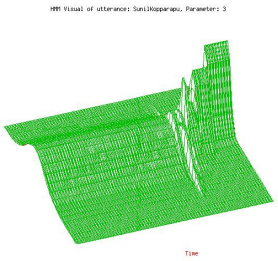
Figure 2: Visual representation of HMM corresponding to the first 8 parameters of the spoken word *Appointment*.



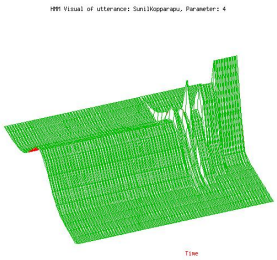
(a) Par 1



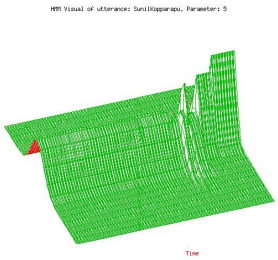
(b) Par 2



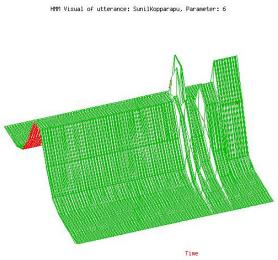
(c) Par 3



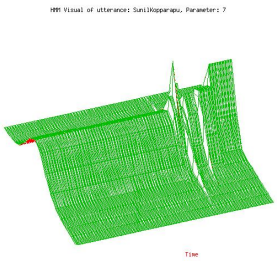
(d) Par 4



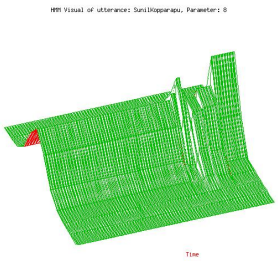
(e) Par 5



(f) Par 6



(g) Par 7



(h) Par 8

Figure 3: Visual representation of HMMs corresponding to the first 8 parameters for the spoken word *Sunil Kopparapu*.

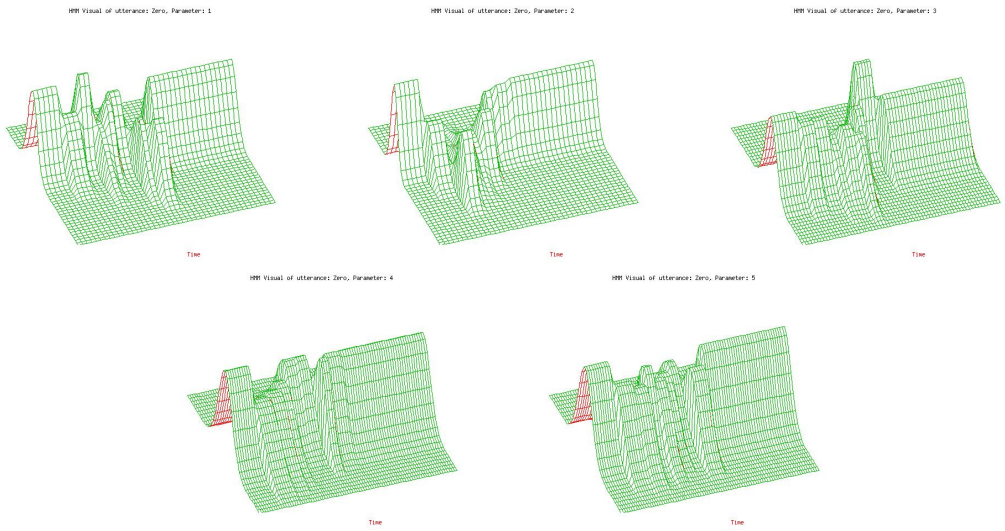


Figure 4: Digit 0

It is a well known fact that, short words pose recognition problems, the classic example is that of being able to recognize the nine spoken digits in English (*one, two, three, four, five, six, seven, eight* and *nine*). Figures 4-13 show the visual of the first 5-MFCC coefficients of the spoken digits *zero* to *nine*. **Some comments on confusion between digits to come here**

3 Conclusions

Hidden Markov models are in use in several diverse areas. HMMs have been used for the last couple of decades especially in speech recognition and continue to be used even today. While they are important, very little is found in literature which can give some insight into *what HMM is really like*. In this paper we have motivated the need for a tool to capture HMM visually. We have used public domain tools to present HMM visually. The developed 3D tool captures all the characteristics associated with an HMM, namely, (i) number of states, (ii) transition probability matrix and (iii) Gaussian distribution corresponding to each state.

References

[1] Y. He and A. Kundu, "2-D shape classification using HMM," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1172–1184, 1991.

[2] N. Kamath, K. S. Kumar, U. B. Desai, and R. Duggad, "Joint segmentation and image interpretation using HMM," in *Proceedings of the International Conference on Pattern Recognition*, (Brisbane, Austalia), August 1998.

[3] J.-L. Chen and A. Kundu, "Unsupervised texture segmentation using multichannel decomposition and HMMs," *IEEE Transactions on Image Processing*, pp. 603–619, 1995.

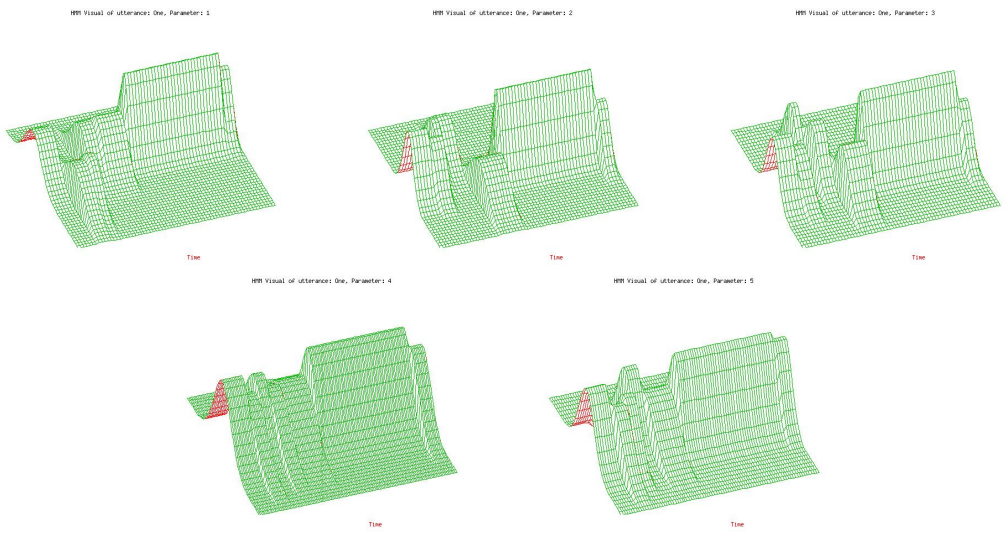


Figure 5: One

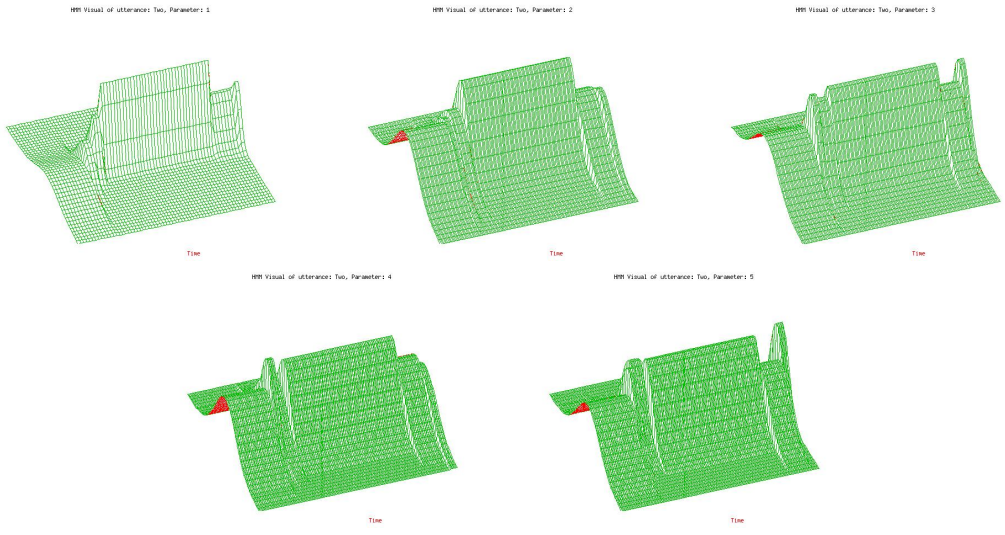


Figure 6: Two

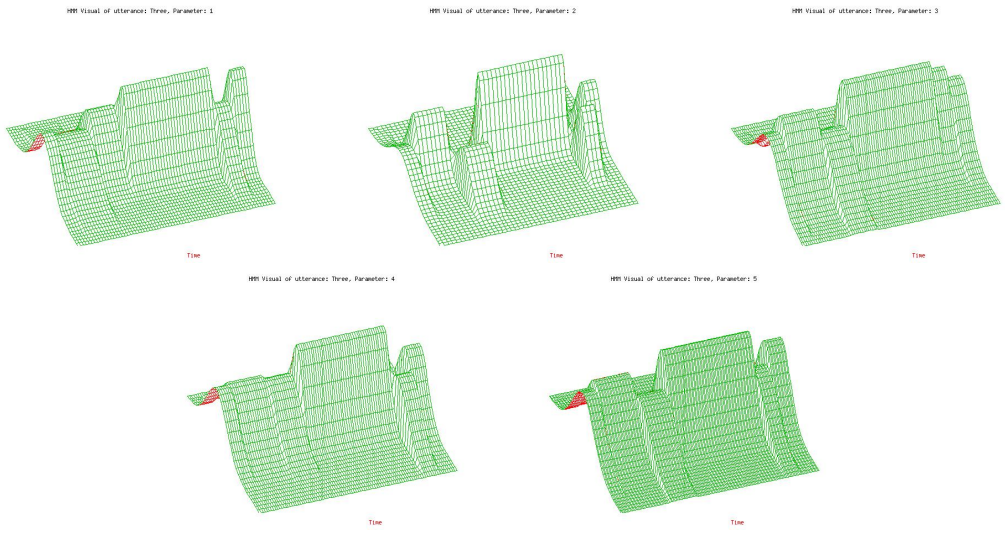


Figure 7: Three

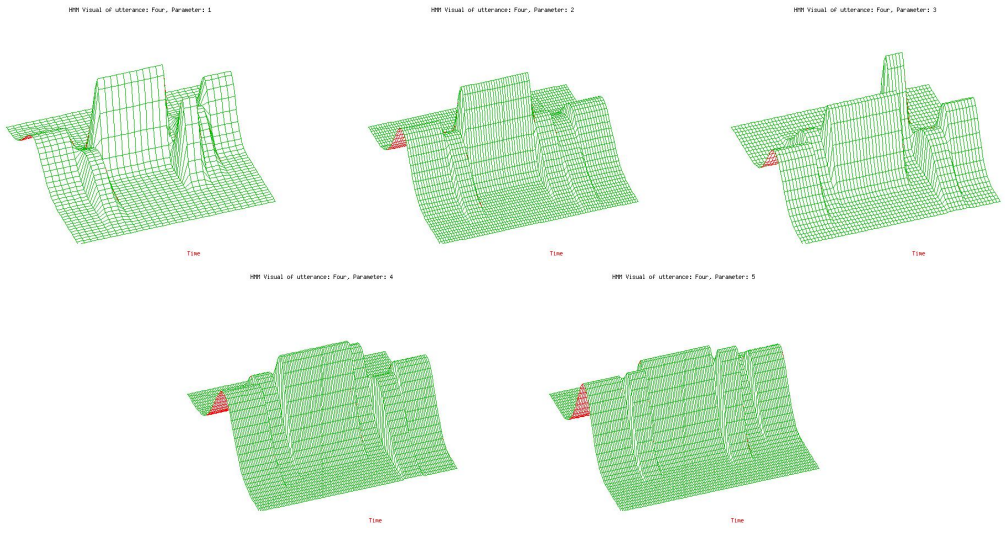


Figure 8: Four

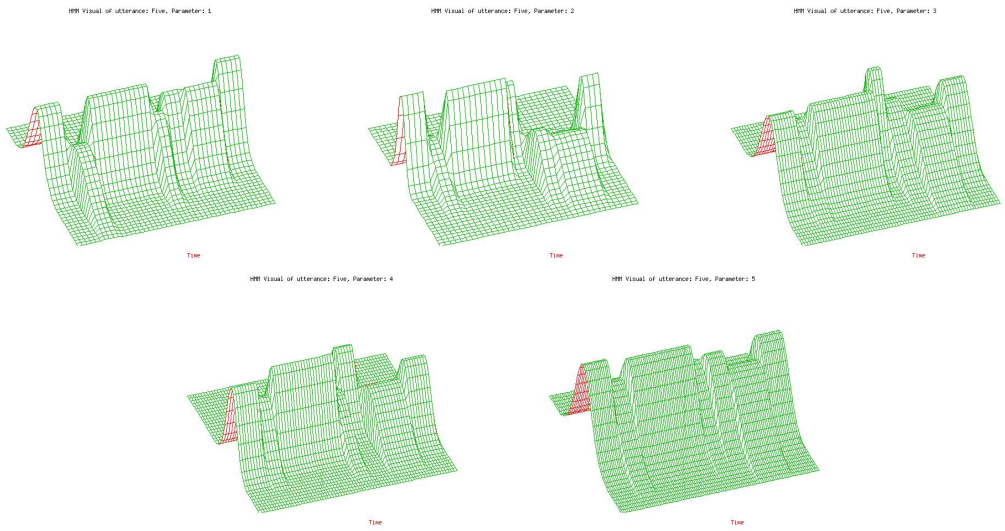


Figure 9: Five

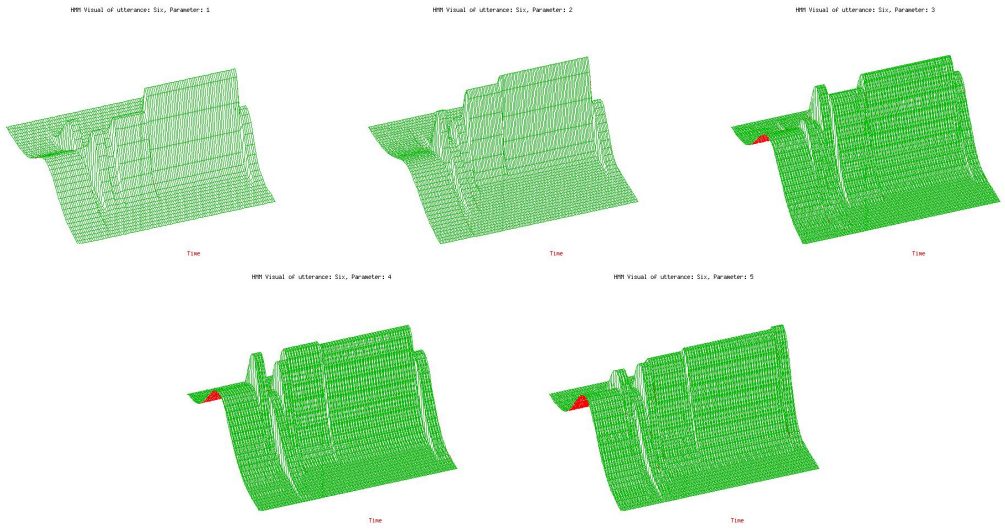


Figure 10: Six

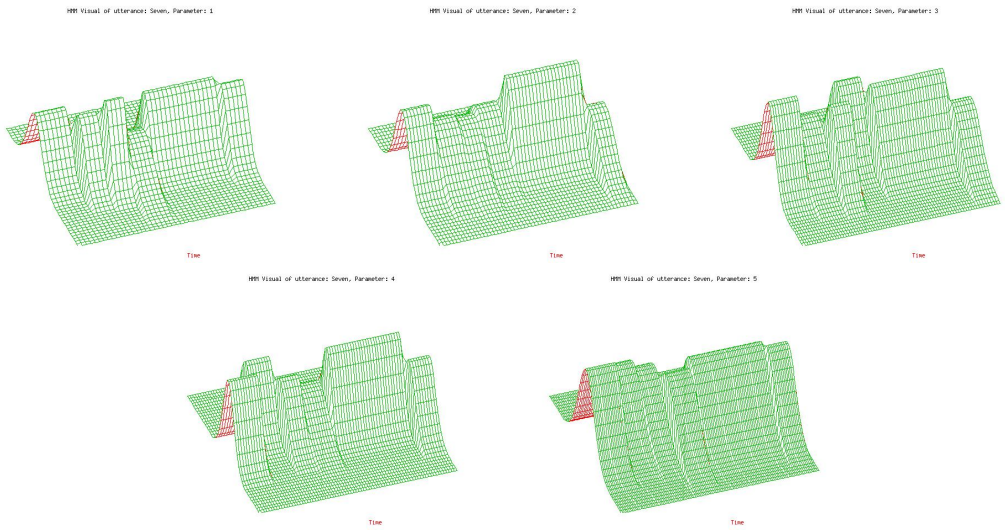


Figure 11: Seven

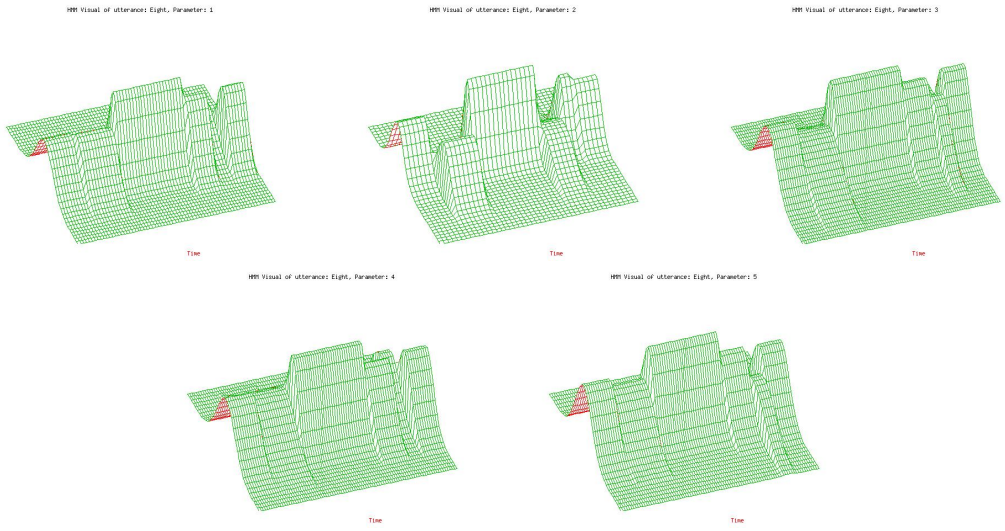


Figure 12: Eight

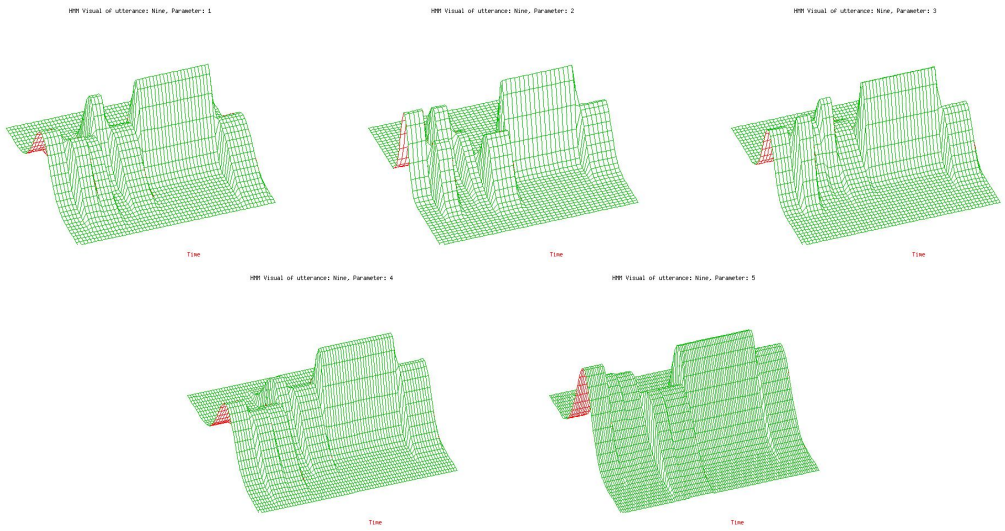


Figure 13: Nine

[4] E. Charniak, *Statistical Language Learning*. Cambridge, Massachusetts: MIT Press, 1993.

[5] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–63, 1998.

[6] C. Lee and Y. Xu, "Online, interactive learning of gestures for human/robot interfaces," *IEEE International Conference on Robotics and Automation*, vol. 4, pp. 2982–2987, 1996.

[7] 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, 1989.

[8] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, pp. 4–16, June 1986.