Arousal Detection in Neonatal EEG Signals Using Hidden Markov Models

Senior Project Final Report

Randall S. Kelso

August 2002

Advisors:
Ken Loparo
John Turnbull

EXECUTIVE SUMMARY

The goal of this project was to study methods of using Hidden Markov Models (HMMs) to automatically detect sleep arousals in the electroencephalogram (EEG) signals of neonates. A sleep arousal is characterized as a frequency shift of energy in the EEG during sleep involving a change of stae in the neonate. These changes can be very difficult to identify due the extreme variability and overlap of signal statistics in arousals and non-arousals, as well as from patient to patient. HMMs have become increasingly popular as a classification technique due to their success when applied to problems in the field of speech recognition.

Our method of detecting arousals using HMMs has proved to give excellent results in most states of sleep. More work needs to be done developing a different feature extraction technique to improve the results of this method in the remaining state of sleep.

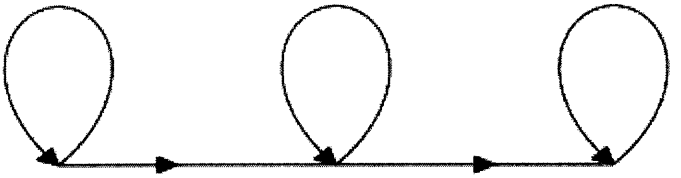
INTRODUCTION

This project was inspired by the research of Mark Scher, a member of the Neurology department at University Hospitals. Dr. Scher requested that an automated arousal detector be developed here in the EECS depart at CWRU. The CWRU engineering group recommended trying to apply HMMs to this problem. Dr. Scher, and other neurologists, spend many hours reviewing Polysomnographic (PSG) data, including EEG data, noting many important events, including arousals, when diagnosing the health of premature babies. Understanding the frequency and duration of arousals is an important step to understanding the health of a patient's sleep, especially in neonates as they are in a critical stage of brain development. An automatic arousal detector would help physicians save time diagnosing these patients. Dr. Scher's theory suggests that an abnormally high frequency of arousals constitutes a higher risk of significant problems in brain development.

A PSG arousal can be defined as "an abrupt shift in EEG frequency," but is subject to many subjective rules and exceptions. Often, low frequency movement artifacts are present during an arousal as well as an increase in electromyographic (EMG) amplitude [1]. Physicians typically also look at other information besides in the PSG besides EEG data such as heart rate, breathing rate, visual cues, and state of sleep. There are four states of sleep for a healthy infant: 1) Mixed Signal (MAS), 2) Tracé Alternat (TA), 3) High Voltage Slow (HVS), and 4) Low Voltage Irregular (LVI). Examples of typical arousals are shown in Appendix A.

HMMs were first developed as a tool for automated speech recognition in the late 1960's [2]-[14] and were later applied to many other applications such as detecting tool wear and biomedical engineering [15]-[16]. They are based on discrete Markov processes. These processes consist of several states that can represent the output of certain random variables. A set of transition probabilities connects these states by describing the likelihood that the model will move to the next. Fig. 1 shows an example state topology and the possible state transitions. The Hidden Markov Model is an extension of this Markov process as the actual sequence of states is hidden to the observer.

Figure 1:



As shown in reference [2], an HMM is completely characterized by the following parameters:

- N, the number of states in the model.
- M, the number of distinct observation symbols per state.(number of probability density functions per state?)
- A, the state transition probability matrix. This is an N x N matrix.
- B, the observation symbol probability distribution matrix. There is a matrix b_j for each sate j and is an N x M matrix.
- Q, the set of states. q_t represents the current state.
- π, the initial state distribution. This is a vector of length N.
- O_t, an observation at time t.

b_j is approximated by a weighted sum of M gaussian distributions as

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm}\eta(\mu_{jm}, U_{jm}, O_t) \tag{1}$$

where c_{jm} is a weighting coeffcient, μ_{jm} is the mean vector, U_{jm} is the covariance matrix, and $\eta(\mu,U,O)$ is the multivariate Gaussian probability density function given by:

$$\eta(\mu,U,O) = \frac{1}{\sqrt{(2\pi)^n \mid U \mid}} \exp(-\frac{1}{2}(O-\mu)'U^{-1}(O-\mu)) . \tag{2}$$

Thus the model λ can be represented as:

$$\lambda = (A,B,\pi) = (A, c_{jm}, \mu_{jm}, U_{jm}, \pi) . \tag{3}$$

The models output a probability which represents the likelihood that the given observation is a member of the class that the model was trained to recognize. This probability is given by:

$$P(O \mid \lambda) = \sum_{i=1}^{N} P(O, q_t = i \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \tag{4}$$

where α and β are the forward and backward variables. α is given by:

$$\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i \mid \lambda) \tag{5}$$

References [2] and [15] provide extensive examples of these calculations. Please refer to references to them for more complete details.

In this project, we used the HMM Toolbox for Matlab written by Hasan Ocak to determine these parameters and evaluate the outputs of the HMMs.

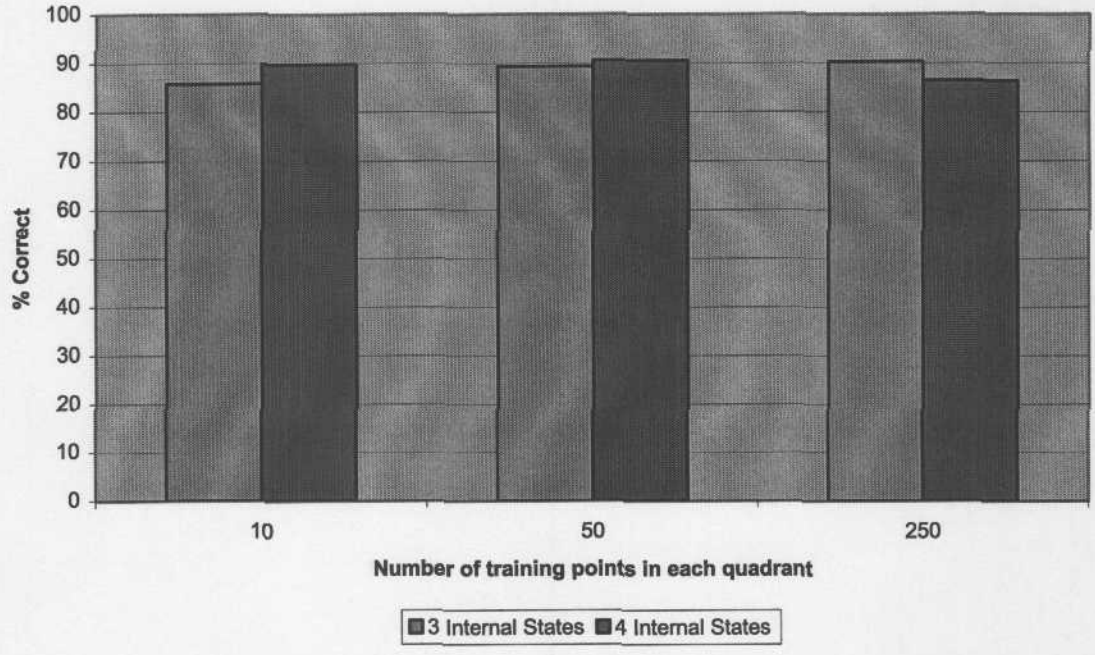
The parameters are estimated by supervised training. We give the model a large set of observation points from on class of data. Then, we repeat the process with observations from another class of data for as many classes as we are interested in. For this project, we trained the model on two sets: arousals, and non-arousals.

METHODOLGY

One of the most valuables lessons I learned during this project is that one will almost never solve a large, complex problem on the first try. It is much more productive to systematically solve the problem step by step. Each block must be designed, tested, and proved to be working reliably before one can expect the second block to work properly based on the output of the first block. Because of this fact, we chose to study much simpler, controlled problems before attacking the PSG data.

The first step was to understand how HMMs work and how to implement them with Matlab code we had available. We decided to create points randomly distributed along the unit circle and then classify these points as to which of the four quadrants these points lie in. First, we generated a random angle uniformly distributed between 0 and 2π and trained an HMM to classify which quadrant this angle lied in. This model contained only one feature. After we had this model working, we doubled the complexity of the system by classifying randomly generating x,y pairs on the unit circle. We adjusted the parameters of the HMM to try and improve the accuracy. We varied the number of states, the number of density functions per state, the number of training points, and the maximum and minimum allowed variance value. The results of these experiments are shown in figure 2.

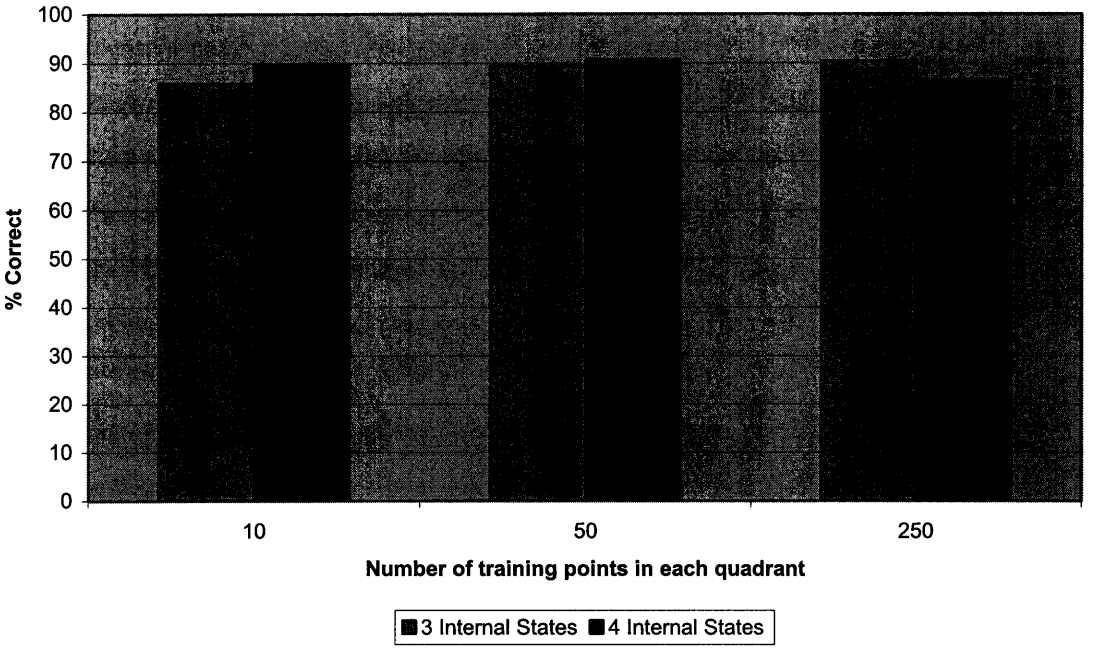
Fig. 2: Accuracy of HMM, Quadrant Classification Problem



Next we wanted to classify signals similar to PSG signals, but in a controlled environment. We created a set of sine waves with a random frequency and then did feature extraction on these signals using the same method we planned on using with the actual PSG data. We took the Fast Fourier Transform (FFT) of the signal and found the total power in each of the four frequency bins we created. This classification method proved to be almost perfect, correctly classifying the signal over %99 of the time.

After we achieved success classyifying the frequency of the sine waves, we were curious to see how the HMMs responded to errors in the training sets. This was a necessary step as an effective arousal detector would have to be trained on a large set of data likely to contain many misclassified arousals due to the variable nature of an arousal. We added gaussian noise to the frequency component of the signals causing some signals to be placed in the wrong training sets. The preliminary experiments show that, given enough training points, the HMMs could still perform remarkably well even with bad training data. Figure 3 shows the results of these experiments. There were no errors contained in the testing set.

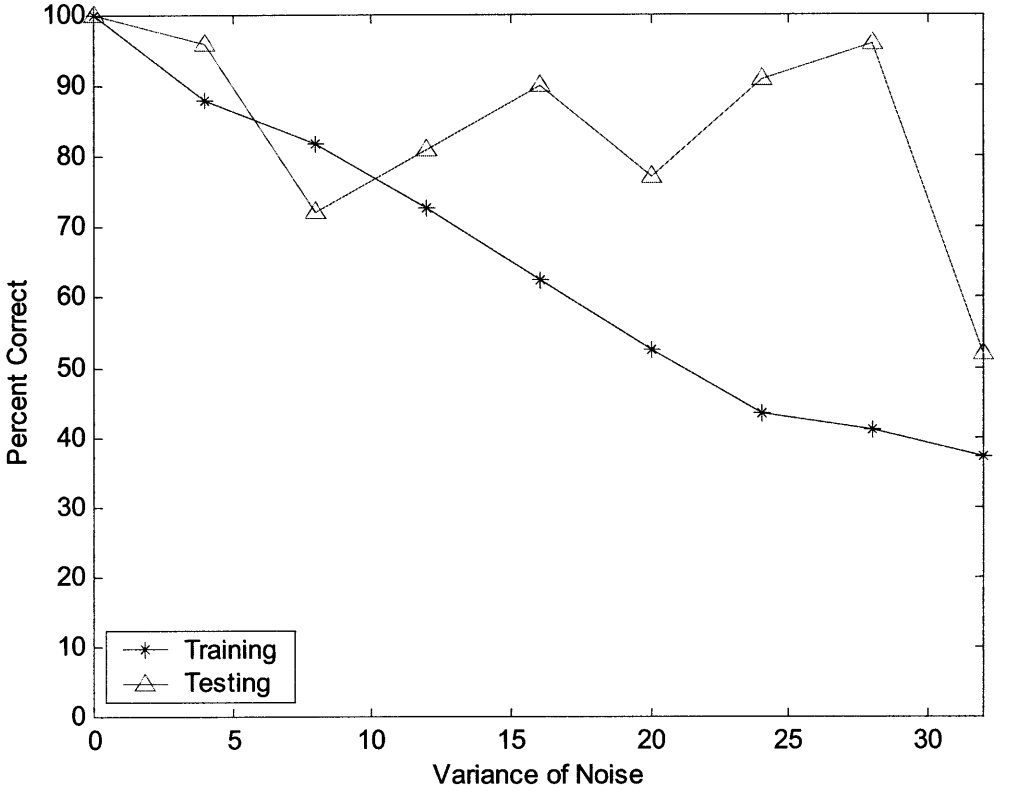
Fig. 2: Accuracy of HMM, Quadrant Classification Problem



Next we wanted to classify signals similar to PSG signals, but in a controlled environment. We created a set of sine waves with a random frequency and then did feature extraction on these signals using the same method we planned on using with the actual PSG data. We took the Fast Fourier Transform (FFT) of the signal and found the total power in each of the four frequency bins we created. This classification method proved to be almost perfect, correctly classifying the signal over %99 of the time.

After we achieved success classyifying the frequency of the sine waves, we were curious to see how the HMMs responded to errors in the training sets. This was a necessary step as an effective arousal detector would have to be trained on a large set of data likely to contain many misclassified arousals due to the variable nature of an arousal. We added gaussian noise to the frequency component of the signals causing some signals to be placed in the wrong training sets. The preliminary experiments show that, given enough training points, the HMMs could still perform remarkably well even with bad training data. Figure 3 shows the results of these experiments. There were no errors contained in the testing set.

Fig. 3: Percent Correct Vs. Variance of Noise in Training Set



Now that we had a method that works robustly in the ideal world, we began to work with real PSG data. First, we adapted code John Turnbull had used earlier to import data from the linux server and convert it from multiplexed 16-bit binary files into a data format Matlab could use. The script creates a matrix that has a row for each channel of EEG data and a column for every sample of data. It also reads the sample rate and the annotations the neurologist added to the data file. These include the start and stop times of arousals, the state of sleep, and many other comments. This function extracts just the pertinent arousal information. Next, we wrote a function that sorts one channel of data into a vector of arousal data and a vector of non-arousal data based on the comments made by the neurologist.

Feature Extraction

The next stage of the development process, finding a working feature extraction technique and debugging the entire process, took considerably longer than any other. We wrote another function that takes the FFT of one second epochs of each channel. The data is zero padded to get 0.25 Hz resolution. Before taking the log of the FFT, the function sets any values of 0 to the smallest value on the computer (about 10^{-300}) so as not to produce any negative floating point underflows. We take the log to magnify the small changes that take place at the higher frequencies. This is necessary because otherwise the model would react only to low frequencies changes where the majority of the energy is. The log adds more weight to the high frequency changes in the HMM's computations.

Next, we sum all of the energy in each of 5 frequency bins and divide by the width of each bin. These frequency bins are 1) Delta (< 4 Hz), 2) Theta (4 to 8 Hz), 3) Alpha (8 to 14 Hz), 4) Beta (14 to 24 Hz), and 5) EMG (24 to Nyquist, which is 32). We later realized that taking the log before we summed the energy was like multiplying each point in the FFT and then taking the log. This does not produce features that are the power in each of the frequency bins, however, we later found it worked better than the expected method of taking the sum before the log. We also found later that using log base ten produces better results than using natural log. We also tried using spectral smoothing windows, such as the Hanning window, but these were abandoned as they increased the frequency of false positives.

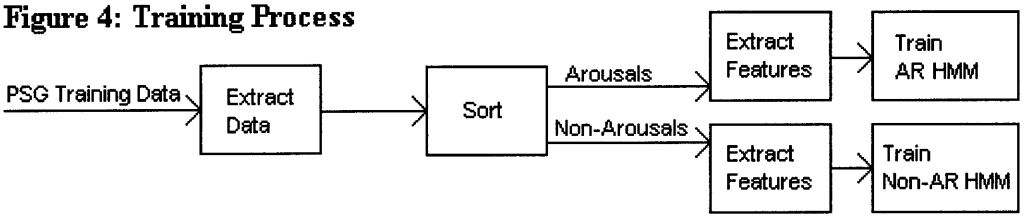
Another method we experimented with involved doubling the number of features vectors by inputting a second point of data five seconds behind the first set. We thought giving the model the ability to see what had just occurred in the PSGs would help. This is information the physicians can use to help them diagnose the neonates. However, this confused the HMM. It may be worthwhile to look into this method further, but using four classes of HMMs: 1) both windows are not in an arousal, 2) the leading window is in an arousal, 2) both windows are in the arousal, and 4) the leading window only is not in an arousal event.

Training

The training process usually takes hours of computation time. This proved to be quite a hindrance. We found we could reduce this time by carefully selecting only the four most pertinent PSG channels and using only three states with two probability density functions per state without a significant loss of accuracy. We carefully chose four of the 14 channels we had available to provide good coverage of the entire brain. These four channels were Fp_1-T_3, Fp_2-T_4, C_3-O_1, C_4-O_2. We also tried using the supercomputing center at OSU but found that the desktop machines were faster.

To classify the arousals we actually build two HMMs: one for arousals, and one for non-arousals. One model is trained on data from arousals, while the other is trained on the rest of the data. The training process is outlined in Fig. 4.

Figure 4: Training Process



Classification

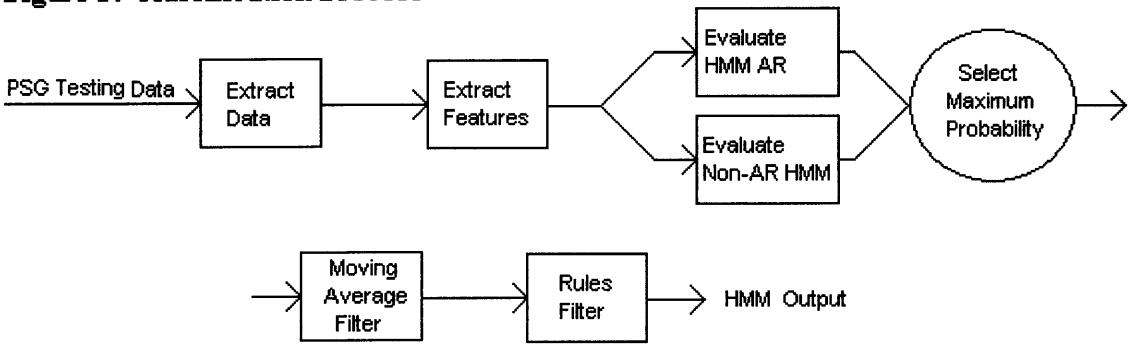
Now that the model is trained we enter the classification phase. To classify a section of PSG data we extract features exactly as we did to train the model. We then pass these features to the HMM and evaluate the probability that the given observation is a member of the arousal class and evaluate the probability that the given observation is a member of the non-arousal class. Whichever model outputs the maximum probability determines whether or not an arousal is occurring for the given observation.

Next, we designed two filters to clean up the output signal of the HMMs. The first is a simple five-point moving average filter to smooth the signal. It removes transients in the output stream. This is to ignore 1 or 2 second changes in the state of the HMM. Next, the output of the moving average filter is passed to a "rules" filter. This checks to make sure each arousal is at least a minimum length of time long, and that the time inbetween

arousals is also at least a minimum length of time. Mark Scher initially suggested 5 and 10 seconds respectively for these lengths.

After experimentation was complete, two final models were built. One was trained on various pieces of data from one patient, while the other was trained on pieces of data from five separate patients. These patients were all less than six months old. The classification process is shown in the block diagram in Fig. 5.

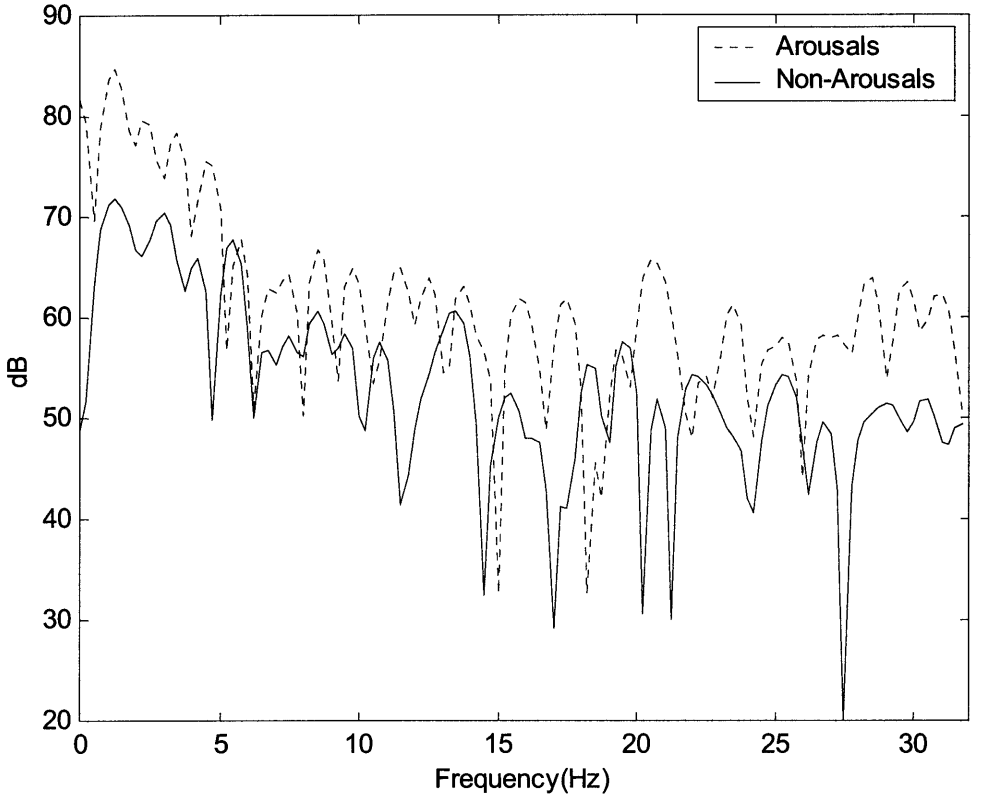
Figure 5: Classification Process



RESULTS

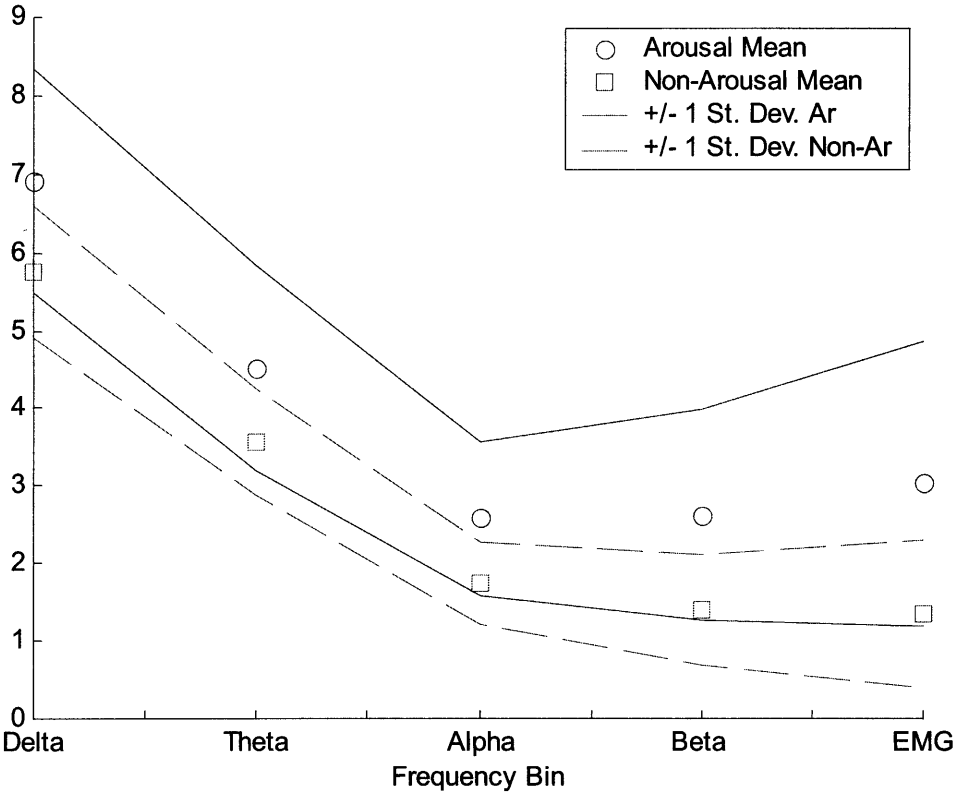
Figure 6 shows the average Log Power Spectral Density of arousals and non-arousals. A Hanning window was used to smooth the PSD.

Figure 6: Log Power Spectral Density



It can be seen that there is slightly more energy in the arousals across the band. Figure 7 shows the statistics of the frequency bin features. It is clear that the mean energy of the arousals is greater, but there is a significant overlap in the standard. Not all arousals have more energy than non-arousals and mean energy cannot be used as a single feature alone. These statistics vary greatly across the four different states of sleep.

Figure 7: Frequency Bin Statistics



The final model trained on a single patient, classified %72.06 of the observations correctly in the training set. A more complete description, including false-negatives and false-positives, is shown in Table 1. It should be noted that it performed very well in three states of sleep, with the Tracé Alternant state pulling the average down. As expected, this model performed less well on another patient (%64).

Table 1

		Actual	
		Arousal	Not-Arousal
Detected	Arousal	20.56	24.50
	Not-Arousal	3.00	51.50

The second final model to be built classified %87.36 of the total number of observations in the training set correctly. In the testing set, the model classified %83.43 of the observations correctly with a maximum of %89 and minimum of %74. The complete contingency table is shown in Table 2. The testing set was comprised of five other patients in the same age group (0 to 4 months) as the training group. Experiments on patients older than this age group received scores of less than %60.

Table 2

		Actual	
		Arousal	Not-Arousal
Detected	Arousal	23.31	6.56
	Not-Arousal	9.43	60.12

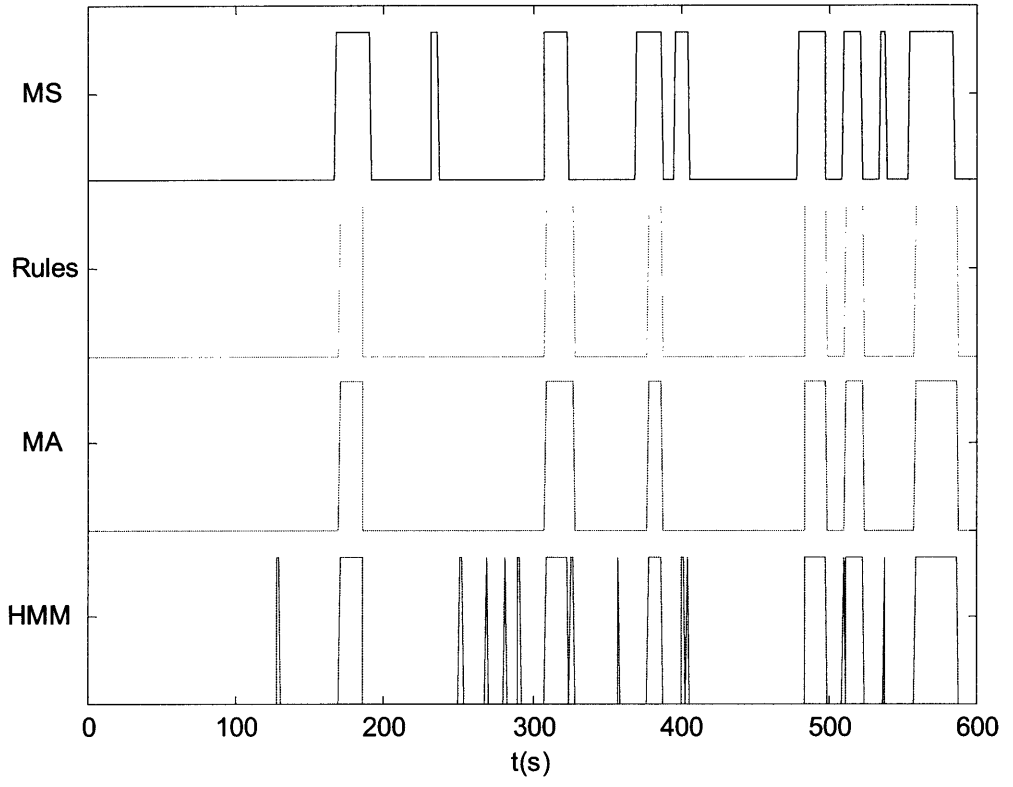
Table 3 gives the results of the automatic detection algorithm for each of the two final models broken down by state of sleep.

Table 3

		Trained on 1 patient		Trained on 5 patients	
		Actual		Actual	
State of Sleep:	Detected:	Arousal	Not-Arousal	Arousal	Not-Arousal
MAS	Arousal	12.34	1.16	9.47	0.28
	Not-Arousal	6.17	79.69	10.59	78.97
HVS	Arousal	0.00	5.04	0.00	0.00
	Not-Arousal	4.62	88.24	4.62	93.28
TA	Arousal	5.84	35.54	1.25	0.00
	Not-Arousal	1.60	56.68	6.19	92.21
LVI	Arousal	61.10	17.99	52.30	5.65
	Not-Arousal	3.56	16.32	12.34	28.66

Figure 8 is a sample graphical representation of the outputs of the HMM, the Moving Average filter (MA), and the Rules filter as compared to classifications done by the clinician, Mark Scher (MS)—the "desired" results. The "high" level represents arousals while the "low" level represents non-arousals. In this ten-minute example our method correctly identified six of the nine arousals with zero false negatives. It classified each second %89.96 correctly with %1.5 false positives and %7.69 false negatives. This is from the MAS state of sleep.

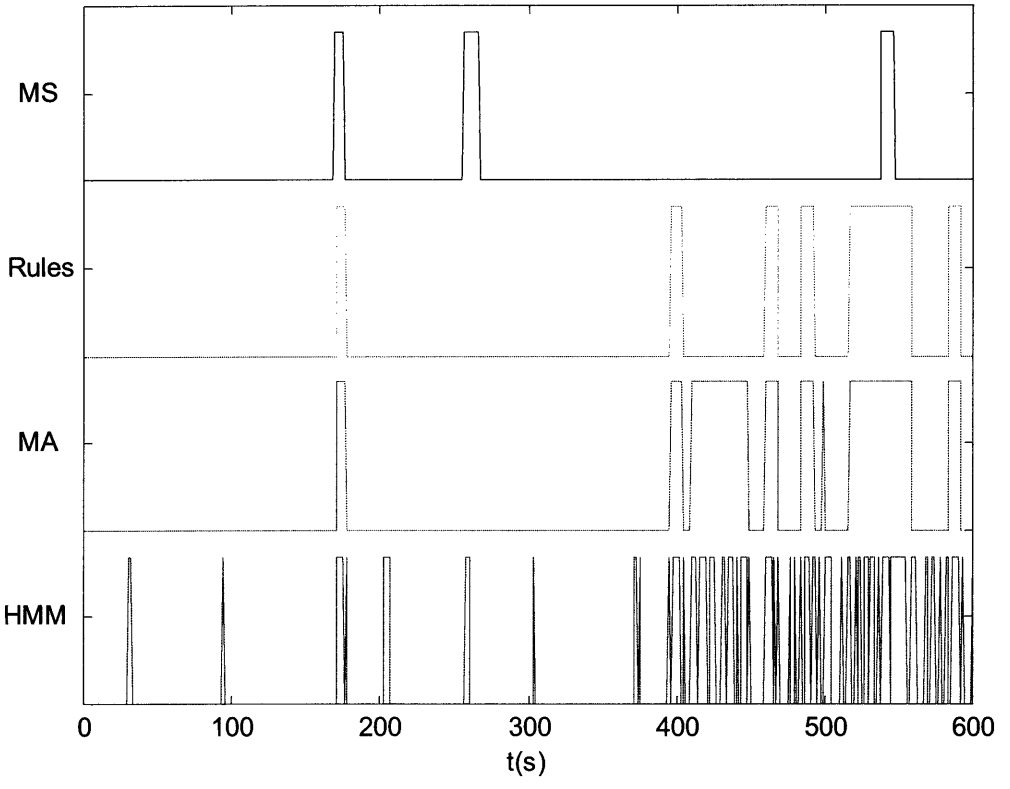
Figure 8



CONCLUSIONS

As the data shows, our project made extensive progress in designing an automatic arousal detector. From Table 3 and Fig. 8, one can clearly see our classification algorithm performs excellently in the MAS state of sleep. The algorithm performs well in the LVI state of sleep. While the data shown in table 3 for HVS is modest, I am not concerned because this breaks down to 0.25 false negatives per minute of data. HVS is a more rare state of sleep at this age group. Also, the MAS, LVI, and HVS states of sleep are relatively stationary with respect to arousal detection.

Figure 9: Minutes 10-20, trained on 1 patient



However, HMM fails on Tracé Alternant when using the feature extraction techniques implemented in this project. In Fig. 9, the presence of the rapidly changing square wave signifies when the patient is in the TA state of sleep. Here the HMM performs poorly, thinking the patient is rapidly moving in and out of arousal states. This is due to the fact

that Tracé Alternant sleep contains rapid short bursts that look similar to arousals. The rules for classifying arousals are different in each state, especially in Tracé Alternant.

Fig. 10 shows the same input section of data as Fig. 9 but the output is from the model trained from five different patients. These results are the opposite of the first model: the HMM did not detect any arousals in this state.

Figure 10: Minutes 10-20, trained on 5 patients

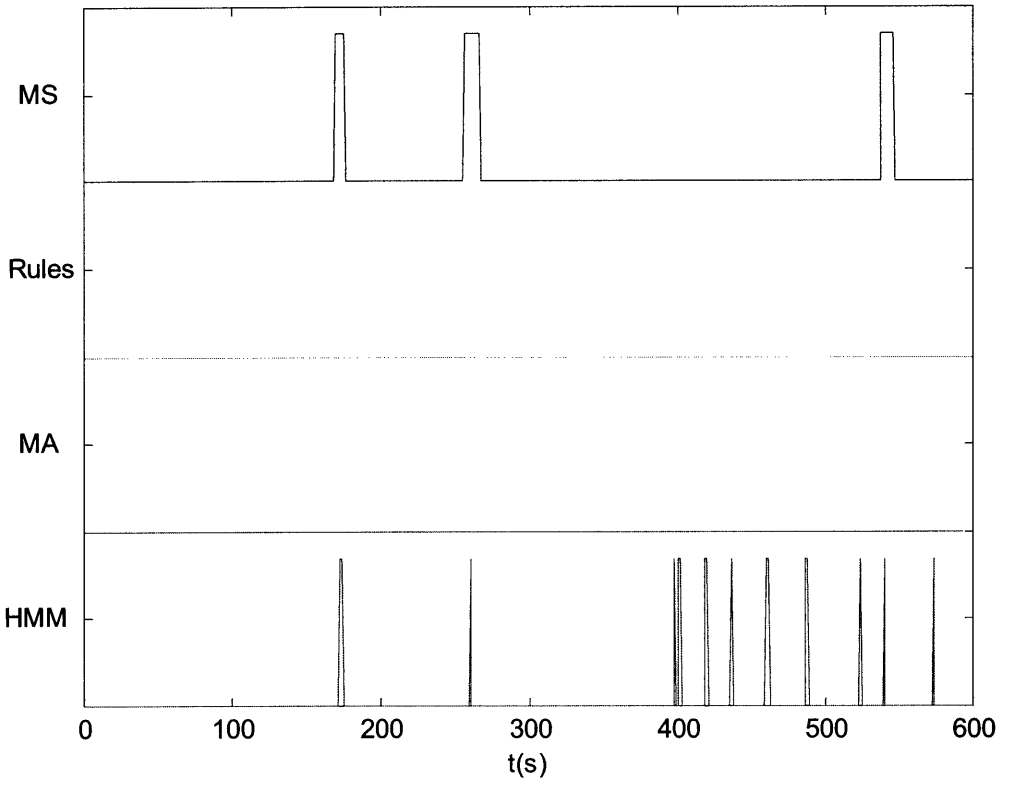
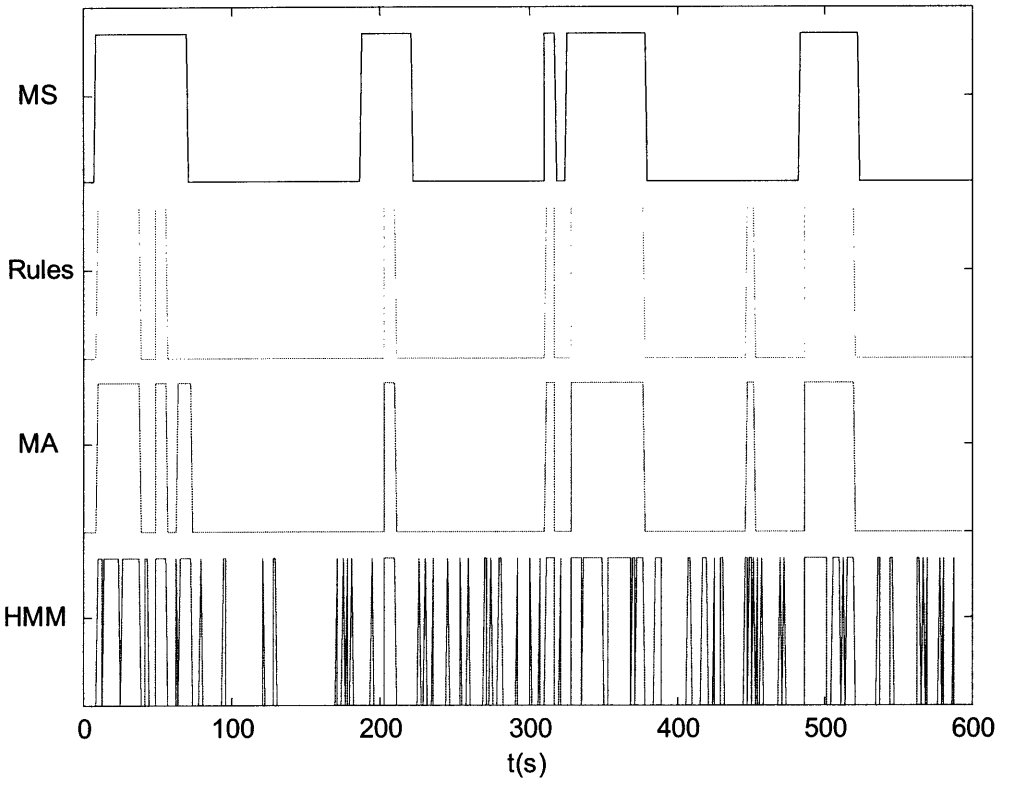


Fig. 11 shows the benefits of the Moving Average and Rules filters. Here the Moving Average filter does an excellent job removing the pertinent information from a less than perfect HMM output. It not only removes any short false positives, but also groups together large sections of correctly identified arousals by removing short transitions back to the non-arousal state. The Rules filter removes one of the extra arousals during the first 100 seconds of the figure. This brings the total detected arousals down to seven from eight in a section of data where the correct number of arousals is 5. Obviously, two of these are contained in one of the actual arousals. The Rules filter does a better job and detecting the correct number of arousals while the Moving Average filter classifies each second with more accuracy.

Figure 11

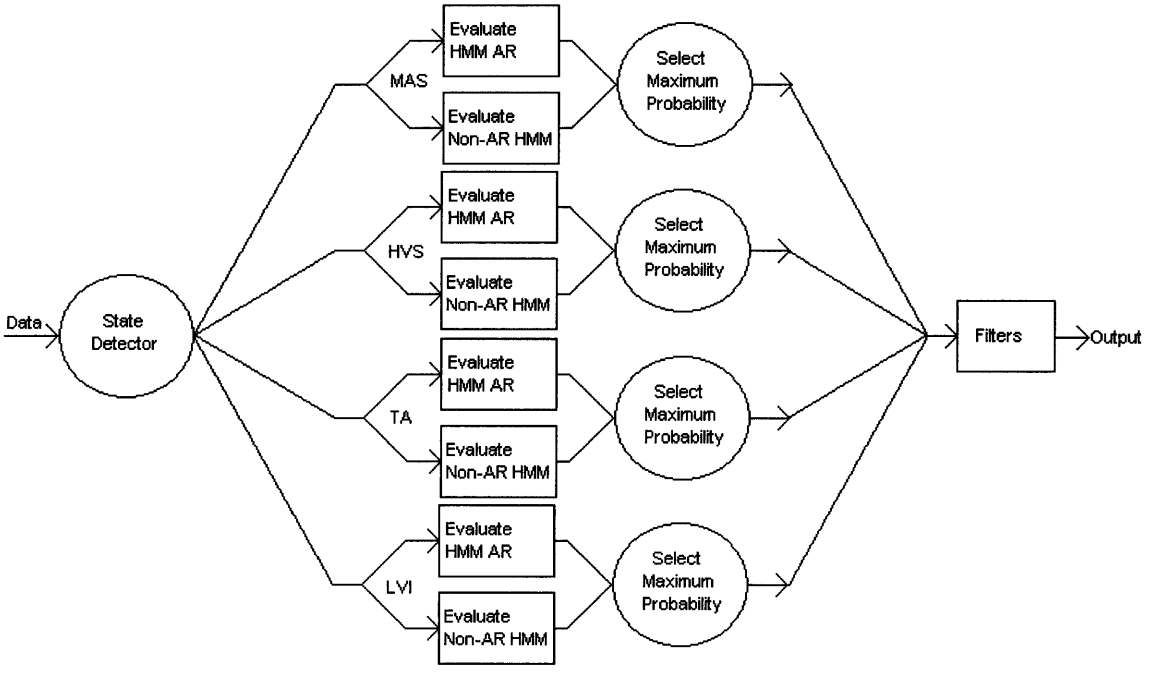


The false positive arousal occurring at about 450 seconds brings up an interesting point. This may be an actual arousal that the clinician missed in his first inspection of this patient's PSG. Due to the subjective nature of these arousals, it is possible that if the same clinician were to rescore this patient, he would count this arousal.

RECOMMENDATIONS

To further improve the performance of this method, a more complicated approach is proposed. Build four separate HMMs, one trained for each state of sleep. Different feature extraction methods could be used for each stately. Other methods could include autoregression (AR), linear prediction code (LPC), or wavelets. This will certainly be necessary for Tracé Alternant. The state of sleep will have to be known during the classification state. This could be provided by data from the clinician, but eventually should be automatically detected as well. HMMs would be an excellent method for state decection as well. These models should be trained on a much wider range of patients and could include other information extracted from the PSG, beside EEG, such as EMG, EKG, and breathing. Fig. 12 shows a possible implementation of these recommendations:

Figure 12:



REFERENCES

[1] Atlas Task Force, "EEG Arousals: Scoring Rules and Examples," *Sleep*, vol. 15, no. 2, pp. 174-184, 1992.

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

[3] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554-1563, 1966.

[4] L.E. Baum and J.A. Egon, "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology," *Bull. Amer. Meterol. Soc.*, vol. 73, pp. 360-363, 1967.

[5] L.E. Baum and G.R. Sell, "Growth functions for transformations on manifolds," *Pac. J. Math.*, vol. 27, no. 2, pp. 211-227, 1968.

[6] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximumization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 1, pp. 164-171, 1970.

[7] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.

[8] J.K. Baker, "The dragon system—An overview," IEEE Trans. *Acoust. Speech Signal Processing*, vol. ASSP-23, no. 1, pp. 24-29, Feb. 1975.

[9] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Develop.*, vol. 13, pp. 675-685, 1969.

[10] L.R. Bahl and F. Jelinek, "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognitition," *IEEE Trans. Informat. Theory*, vol. IT-21, pp. 404-411, 1975.

[11] F. Jelinek, L.R. Bahl, and R.L. Mercer, "Design of a linguistic statistical decoder for the recognition of continous speech," *IEEE Trans. Informat. Theory*, vol. IT-21, pp. 250-256, 1975.

[12] F. Jelinek, "Continous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532-536, Apr. 1976.

[13] R. Bakis, "Continuous speech recognition via centi-second acoustic states," in *Proc. ASA Meeting* (Washington, DC), Apr. 1976.

[14] F. Jelinek, L.R. Bahl, and R.L. Mercer, "Continous speech recognition: Statitical methods," in *Handbook of Statistics, II,* P.R. Krishnaiad, Ed. Amersterdam, The Netherlands: North-Holland, 1982.

[15] H.M. Ertunc, K.A. Loparo, and H. Ocak, "Tool Wear Condition Monitoring in Drilling Operations Using Hidden Markov Models," Machine Tools & Manufacture, vol. 41, pp. 1363-1384, 2001.

[16] R.S. Huang, "EEG Pattern Recognition—Arousal States Detection and Classification," Proceedings of IEEE, pp. 641-646, 1996.

[17] D.A. Coast, R.M. Stern, G.C. Cano, and S.A. Briller, "An approach to cardiac arrhythmia analysis using hidden markov models," IEEE Trans. Biomed. Eng., vol. 37, no. 9, Sep. 1990.

[18] J.E. Stockard-Pope, S.S. Werner, and R.G. Bickford, Atlas of Neonatal Electroencephalography, 2nd Ed., Raven Press, 1992.

APPENDIX A

This Polysomnogram shows a very clear arousal due to the very large movement artifacts. The arousal detector easily classified this one correctly.

This arousal is much more difficult to detect due to the lack of movement artifacts.