

# On the Structure & Discovery of Facial Processes: Hidden Markov Models

## Development through Entropy Minimization Algorithms

*Jane H. Maduram*

Ultimately, the aim of science is to understand systems in a micro and macro sense. This broadly involves deciphering the structure of and the relationships between hidden processes and observable events in a given system, which may be done most succinctly by analyzing data sets. As data sets are products of the original process, they indirectly reveal the structural components and processes within the process that contribute to specific outcomes. A mathematical model of a data set must manipulate information in a manner similar to that found in the original process in order to accurately replicate the data set, resulting in an involuntary mirroring of the internal structure of the original process.

An efficient model must, in addition, trim the mechanisms such that no unnecessary commands or steps are made to produce a compact model. The most compact model achievable, according to a theorem in Kolmogorov complexity, will be the best predictor of a complex process,

$$P(A_i | A) = \frac{P(A_i) P(A | A_i)}{\sum_{j=1}^N P(A_j) P(A | A_j)}$$

with probability one.<sup>1</sup> While the task of finding a compact model may seem a relatively casual problem, there is no method for deriving the most compact model possible. The closest approach

available is to search restricted spaces for a compact space via algorithms. This method becomes increasingly limited, however, as the complexity of the model increases.

The face is a case of a complex space that is extremely difficult to model. Facial processes encode speech and emotion, two processes not fully understood by psychologists or physiologists. This adds to the difficulty of determining which mechanisms operate in the face. In order to address

### References

1. Vitányi P & Li M. Minimum description length induction, Bayesianism, and Kolmogorov complexity. Submitted to IEEE Transactions on Information Theory, 1997.
2. Ekman P, Huang T, Sejnowski T, and Hager J. (Eds). Final Report to NSF of the Planning Workshop on Facial Expression Understanding. Technical report. National Science Foundation, Human Interaction Lab., UCSF, 1993
3. Blair C. Interdigitating muscle fibers throughout orbicularis oris: preliminary observations. Journal of Speech and Hearing Research, 1986; 2: 256-6.
4. Blair C & Smith A. EMG recording in human lip muscles: can single muscles be isolated? Journal of Speech and Hearing Research, 1986; 2: 256-6.
5. Gentil M. Variability of motor strategies. Brain Lang (B5H) 1992; 42(1): 30-7.
6. Fisher CG. Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 1968; 11: 796-804.

this difficulty, this preliminary study investigates a novel method of approximating facial processes to examine both the feasibility and success of current models in structure discovery of the face in contrast to an HMM, which gives biological impetus to the mathematical claim supported by Kolmogorov complexity that the most compact model imitates the internal structure of the original process.

$$P_e(\theta) \propto e^{-H(\theta)}$$

## Background

The face contains 44 muscles that play a direct role in displaying emotion.<sup>2</sup> Muscle types present in the face include linear, sheet, and sphincter muscles, each of which respond differently to stimuli. Muscles in the face are highly integrated,<sup>3</sup> making it extremely difficult to accurately stimulate or record from individual muscles through nonintrusive or intrusive means.<sup>4</sup> An additional problem is that the sets of muscles working in conjunction for specific actions may vary from person to person,<sup>5</sup> further complicating the task of determining what simulated muscles should be used for a specific task within a general model.

The elemental mechanics of speech are summarized by phonemes, visemes, and coarticulation. A phoneme is a unit of speech that is audibly unique. A viseme<sup>6</sup> is a unit of speech that is visibly unique. Complications arise when attempting to establish a correspondence between visemes and phonemes. The mapping from phonemes and visemes is many-to-many.<sup>7</sup> This poses problems in systems that simplify the relationship to a one-to-one correspondence such as keyframing,<sup>8</sup> as it ignores one of the most fundamental parts of speech: coarticulation. Coarticulation is the act of physically preparing for a phoneme before or after the word is actually spoken. Visemes produced by phonemes are context-dependent and actively change, depending on the phoneme(s) preceding (carry-over coarticulation) and following (anticipatory coarticulation) it.<sup>9,10,11</sup> The rules defining coarticulation are further complicated by the varying roles of dominance that phonemes possess.<sup>12</sup>

The workings of emotion are difficult to define. Psychological evidence shows that the dynamics and interactions of facial movement are more important than the static expressions themselves.<sup>13,14,15,16</sup> This concept is dramatically shown in the contrast of anger and happiness, where motion is key in distinguishing the two emotions.<sup>17</sup> As with speech, facial movement must be considered within context.

## FACS

The Facial Action Coding System (FACS) has long been the only method available for fragmenting facial expressions.<sup>18</sup> The quantitative aspects of the FACS have evolved in two different directions, each direction coping differently with the same qualitative criterion. One approach effectively bypasses the quantitative nature of most facial animation by focusing on appearances rather than on the mechanics of facial change. The face is simplified to a texture map that may be stretched at action centers, points that the FACS specifies as being primary to movement. The other evolved approach works with the muscles directly connected to these action centers to create a quantitative model. The selected muscles are then added to a model that integrates functions for the bone, muscle, fat, and skin, thus creating a highly realistic and individualistic model that sacrifices a reasonable number of calculations in return for high anatomical realism. In both cases, FACS forces an emphasis on local facial movement.<sup>19</sup> The incompatibility of these two models may reflect in some part the present inability of the computer sciences to satisfactorily resolve the qualitative nature of the FACS.

## Algorithm

In place of the qualitative techniques used by the FACS to approximate facial movements, an algorithm was used during this study to train the Hidden Markov Model, which was then compared against the FACS-dependent model. The model was based on Bayesian probability, a statistical procedure that estimates the parameters of an underlying distribution based on the observed distribution. Bayes's rule is stated in Equation 1. The parameter estimate of Bayesian probability requires the input of a prior distribution, customarily a Gaussian distribution, from which a posterior distribution is derived. The posterior distribution is the normalized product of the likelihood and the prior distribution, and the parameter estimate is the mode of the posterior. While Bayesian analysis is efficient, it is controversial because the validity of the parameter estimate relies entirely on the validity of the prior distribution, which cannot be assessed statistically.

The algorithm used<sup>19</sup> modifies standard Bayesian elements in three integral parts: the prior, a set of maximum a posteriori (MAP) parameters, and trimming functions. While most standard Bayesian priors are typically set to an uninformative value on the premise that the data alone should affect the posterior, the algorithm

7. Owens E & Blazek B. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 1985; 28: 381-393.
8. Parke F. Parameterized modeling for facial animation. *IEEE Computer Graphics and Applications*, 1982; 2(9): 61-68.
9. Henke WL. Dynamic articulatory model of speech production using computer simulation. Ph.D. dissertation, Massachusetts Institute of Technology, 1966.
10. Huffman MK. Patterns of coarticulation in English. *UCLA Working Papers in Phonetics*, 1986; 63: 26-47
11. Whalen DH. Coarticulation is largely planned. *Journal of Phonetics*, 1990; 18: 3-35.
12. Fowler CA. Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, 1981; 46: 127-139.
13. Bassili JN. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology* 1978; 4: 373-379.
14. Bassili JN. Emotion recognition: The role of facial motion and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 1989; 37: 2049-2059.
15. Bruce V. *Recognizing faces*. London: Lawrence Erlbaum Associates, 1988: 1-50.

argues that an uninformative prior underestimates the intrinsic knowledge present of what learning is defined by and of what defines a good model. In order to account for this, an entropic prior is introduced (Equation 2) such that  $P_e$  represents a partial differential function with an entropy term and represents the parameters of the model. This allows the prior to push parameters to strong, information-containing parameters. In contrast to Maximum Entropy methods, which search for the most noncommittal model that fits the data, the entropic prior forces a sparse, structured, and deterministic model for the data.

The maximum a posteriori estimator produces parameters that maximize the probability of the model given the data. The posterior, given the entropic prior, is Equation 3. To determine the MAP of the posterior, the derivative of the

$$P_e(\theta) \propto P(\mathbf{w}|\theta)P_e(\theta)$$

$$\theta_i = \frac{-w_i}{W(-w_i e^{1+\lambda})}$$

log-posterior is set to zero. This ultimately equals Equation 4, where  $W$  is the Lambert  $W$  function.<sup>20</sup> In terms of entropy, the MAP estimator is Equation 5. These estimates drive irrelevant parameters to zero. While it may be viewed as a coarse-grade process-the estimator may overlook patterns in small data sets, when patterns may be viewed as incidental-the estimator is able to reliably determine structure in large data sets by selecting the strongest hypothesis consistent with the data rather than the most unbiased model. Whereas the MAP estimator gradually reduces selected parameters to zero, trimming allows a rapid elimination of those selected parameters. Trimming is typically completed at or near a convergence when a loss in the likelihood is balanced by a gain in the prior. Its purpose is to produce a matrix that consists of independent parameters. In entropic training, parameters are trimmed when a change will increase the entropy faster than the log-likelihood. Trimming operates

$$-\max_{\theta} \log P_e(\theta|\mathbf{w}) = \min_{\theta} H(\theta) + D(\mathbf{w}||\theta) + H(\mathbf{w})$$

as a driving function towards a skeletal model whose surviving parameters become increasingly well-supported and accurate. In honing the matrix, the parameter subspace is transferred to a simpler dimensionality of geometry, where it retrains.

Training is especially useful when used with the Markov Model (MM), where the probability of a future occurrence is not changed by additional knowledge, given the exact state that the model is in.<sup>21</sup> The MM is a mathematical function that defines states and transitional probabilities between those states to model processes. Transitional probabilities determine the path of the model from one state to another, regardless of previous occurrences, allowing the MM to retain context. In irregular and HMM, the latter of which is used in this study, paths between states can be eliminated during training. This allows the removal of states. Following entropic training results in an HMM that is compact and sparse, with an increased capacity to carry contextual information and an increased specificity between states. The model also encourages a dynamically simple model of data in contrast to a statically simple model. Due to its construction, the algorithm functions best as a method of structure discovery; its strengths are prediction and classification.

## Procedure

Landmark features were painted on the face to outline the hairline, eyebrows, nasolabial line, and lips. The data was taken from a single person reciting "The Jabberwocky" by Lewis Carroll in 55 seconds over 1,640 frames. The reading was taped onto a broadcast quality tape at 29.97 fps. The features were then tracked by a program based on Hager's SSD texture-based tracker.<sup>22</sup> Spring tensions were assigned to edges connecting a pair of trackers, thus enabling the program to correct itself when trackers blundered due to rapid movement or temporary invisibility of the landmark feature. Landmark features outlining the three points on the hairline and the bridge of the nose were used to filter out movements of the face and/or the video camera. Visual data consisted of coordinates, each with a corresponding velocity. The velocity was added to preserve the dynamics of facial expression. Sounds were processed by a mix of LPC and RASTA-PLP

16. Bruner JS. & Tauguirri R. The perception of people. In Handbook of Social Psychology. Houston: Addison-Wesley, 1954.
17. Minsky M. The Society of Mind. New York: Simon and Schuster Inc., 1985: 1-336.
18. Ekman P & Friesen W. Manual for the Facial Action Coding System. Palo Alto: Consulting Psychologists Press, 1978.
19. Brand M. Structure learning in conditional probability models via an entropic prior and parameter extinction. Neural Computation (in press) 1998.
20. Essa IA & Pentland AP. Coding, Analysis, Interpretation, and Recognition of Facial Expressions. MIT Media Laboratory Perceptual Computing Section Technical Report No. 325, 1995.
21. Corless RM, Gonnet GH, Hare DEG et al. On the Lambert W function. Advances in Computational Mathematics, 1996; 5: 329-359.
22. Taylor HM & Karlin S. An Introduction to Stochastic Modeling. Boston: Academic Press, 1984.
23. Hager G & Toyama K. The XVision system: A general-purpose substrate for portable real-time vision applications. Computer Vision and Image Understanding, 1997.
24. Hermansky H & Morgan N. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing, 1994; 2(4): 578-589.

features, which is known to be robust to environmental variations.<sup>23</sup> In addition to this, amplitudes, pitch, frequencies, and energy parameters of the first three formants were used.

$$Sx'_n(t) = Sx_n(t) - \frac{\sum_{\tau=1}^T Sx_n(\tau)}{T}$$

$$Sy'_n(t) = Sy_n(t) - \frac{\sum_{\tau=1}^T Sy_n(\tau)}{T}$$

An additional set of recordings were performed on a 12-second recording of the phonemes "ooh," "aah," "shh," and "ssss." Five points were outlined around the lips and on the jaw. Three hundred frames were obtained at 25 frames per second, and energy channels with 9PLP bands were analyzed to provide audible information. HMM models developed with and without the algorithm were trained on the data, then compared. Once the data set was documented, it was subjected to Principal Component Analysis (PCA) to remove redundancies. The remaining dimensions were selected to preserve approximately 90 percent of the variance, simplifying the data to increase the probability of producing a statistically significant HMM.

The data was split into a visual data set and an acoustical data set. The visual HMM was split into two primary components: a finite state machine modeling the dynamics of the face and a set of Gaussian distributions that isolated and associated facial configurations specific to those states. The facial dynamics were integrated with processed modeling from the acoustic data set to create an HMM that responded to vocal and facial input. The animation utilizing the

vocal/facial model uses Candide, a model that contains 100 triangles with most of the action units stipulated by the FACS<sup>24</sup>. It is considered to be a standard in the model-based image coding community.<sup>25</sup> The HMM worked directly with the points, avoiding the action units. The FACS-dependent model relied on linear algebra to manipulate individual action units, approximating the dissection of phrases into phonemes and visemes.

Error estimation was performed in one of two ways. Ambiguity of each state, resulting in a confidence estimate, was calculated by the average possible states per frame. Actual error was calculated using the normalized coordinates  $Sx'_n(t)$ ,  $Sy'_n(t)$ ,  $Rx'_n(t)$ , and  $Ry'_n(t)$ , where these functions are defined by Equations 6 through 9. Here, coordinates  $Sx'_n(t)$  and  $Sy'_n(t)$  represent the coordinates of the training data at the  $n$ th synthesized point of frame  $t$ .  $Rx'_n(t)$  and  $Ry'_n(t)$  represent the coordinates of the trained model, HMM or FACS.  $T$  is the total number of frames. The normalized error for  $p$  points is shown in Equation 10.

## Results

$$Ry'_n(t) = Ry_n(t) - \frac{\sum_{\tau=1}^T Ry_n(\tau)}{T}$$

Of the 42 action units proscribed by the FACS, 11 are defined in CANDIDE. Of these 11 action units, only seven were needed to describe the range of motion present in video. The existing points for CANDIDE did not correlate precisely with the landmark features used, so points were added to precisely match landmark features. These action units were modified so as to be controlled by these added points. The average error derived for the FACS was .6881.

The algorithm-developed HMM trained with five facial points was representative of the larger

$$Rx'_n(t) = Rx_n(t) - \frac{\sum_{\tau=1}^T Rx_n(\tau)}{T}$$

25. Rydfalk M. CANDIDE: A parameterized face. Ph. D. Thesis, Linköping University, Department of Electrical Engineering., October 1987.

26. Li H, Roivainen P & Forchheimer R. 3-D motion estimation in model-based facial image coding. IEEE Transactions in Pattern Analysis and Machine Intelligence, 1993; 15(6): 545-555.

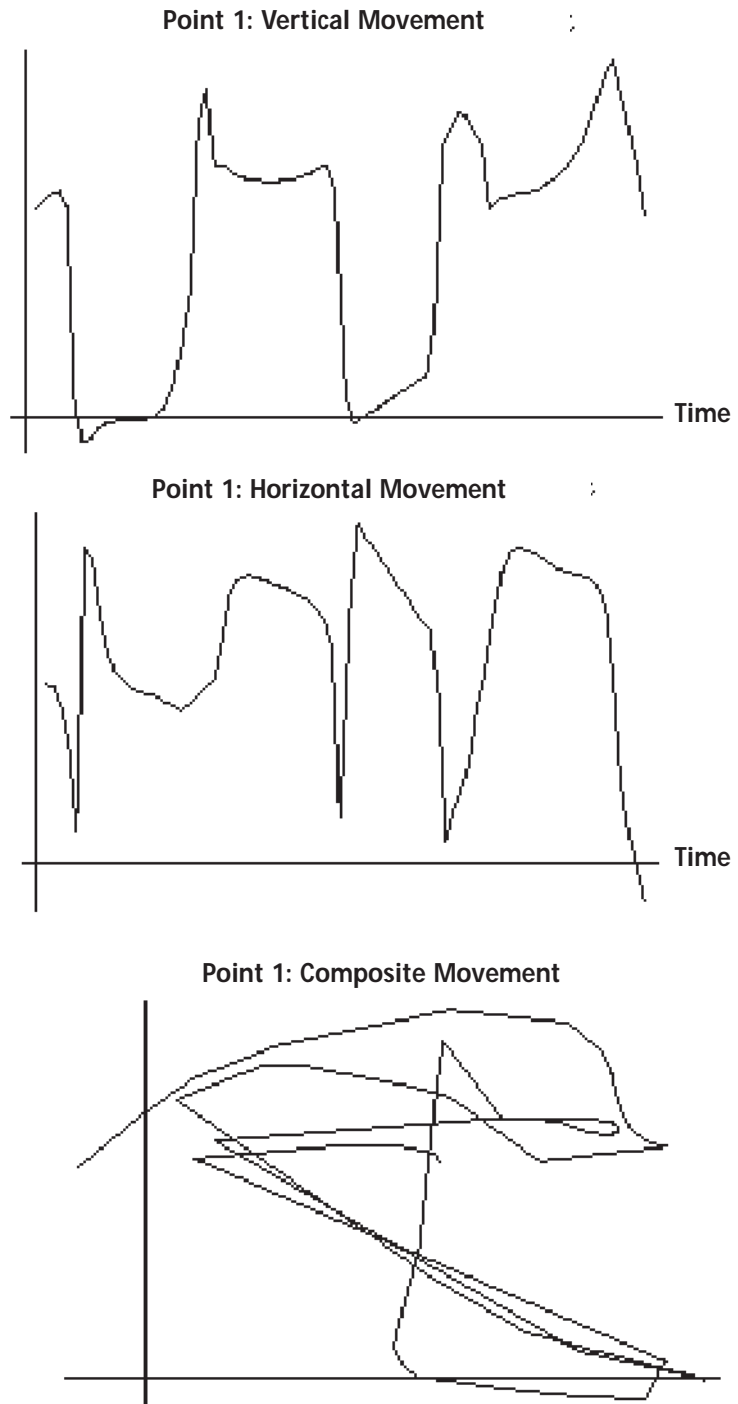
27. Schwartz R, Chow Y, Kimball O et al. Context-dependent modeling for acoustic-phonetic recognition of continuous speech. IEEE International Conference on Acoustics, Speech and Signal Processing, 1985; (): 1569-1572.

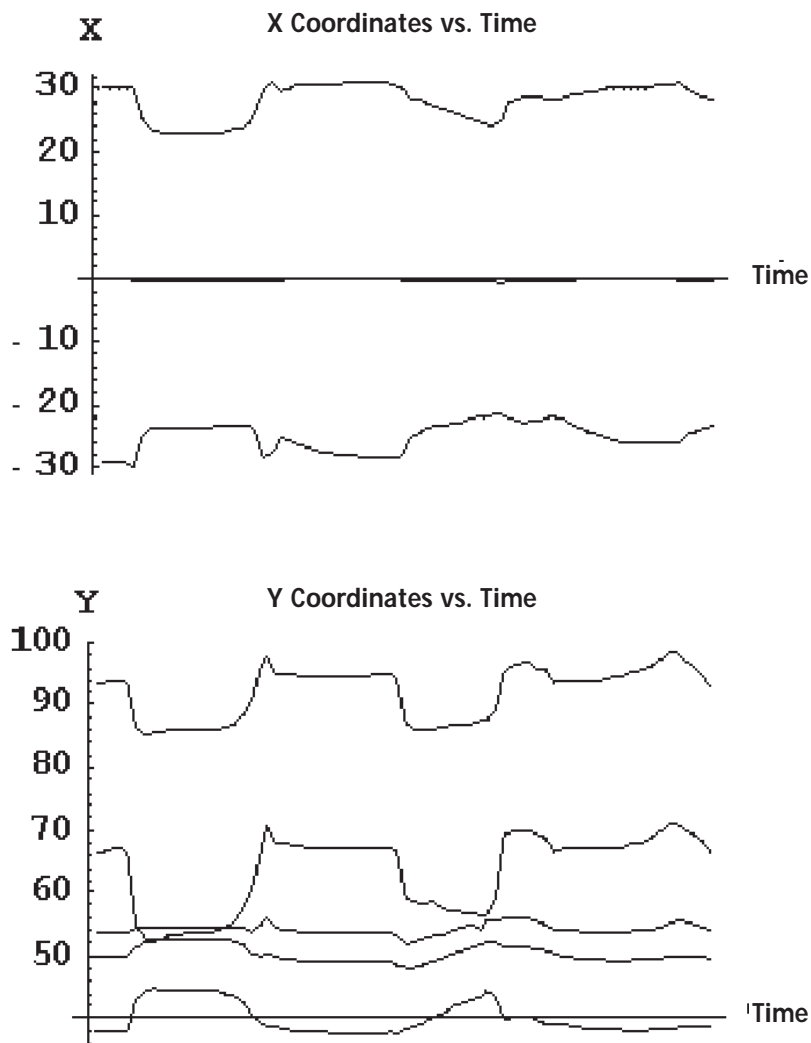
28. Lee KF & Hon HW. Speaker-independent phone recognition using Hidden Markov Models. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989; 37:1641-1646.

$$\frac{\sum_{\tau=1}^T \sqrt{\sum_{i=1}^P \left( (Sx'_i(\tau) - Rx'_i(\tau))^2 + (Sy'_i(\tau) - Ry'_i(\tau))^2 \right)}}{\sum_{\tau=1}^T \sqrt{\sum_{i=1}^P \left( (Sx'_i(\tau) + Rx'_i(\tau))^2 + (Sy'_i(\tau) + Ry'_i(\tau))^2 \right)}}$$

HMMs trained on the entire face. Training data consisted of alternative x and y coordinates of individual points, as demonstrated by the point found on the jaw. This is displayed in Table 1. The movement of the jaw shown in Figure 1 demonstrates the unique properties of the visemes “ooh,” “aah,” “shhh,” and the sibilant “ssss.” Because the 12-second training sequence contained four phonemes, a relatively small 8-state HMM combining auditory and visual information was used to model the 20-dimensional observation vector. The 34 transitions in the sequence produced an average entropy of 0.0204. Each state had an average of one “plausible alternative.” Conventionally trained HMMs produced an average entropy of 1.296, allowing each state in the sequence 1.8 “plausible alternatives.” The visemes clearly visible in the composite graphs of the face (see Figure 2) were effectively synthesized by the algorithm-developed HMM utilizing five landmark features. As Figure 2 demonstrates, horizontal movement of the mouth is more subtle than vertical movement of the mouth and jaw, leading to a necessity of greater detail around the face. This need was addressed in a more extensive series of experiments. (Figure 1, showing movement of the jaw separated into horizontal and vertical components; Figure 2, showing the horizontal and vertical movement of the points surrounding the lips and the jaw).

HMMs integrating 21 landmark features of the face adopted a total of 315 HMMs to the upper and lower parts as well as the entirety of the face to investigate the efficiency of face-to-voice training. The number of states used over these trials varied between 18 and 24 states. Training was completed for 105 different HMMs on the upper face. The most efficient HMM developed contained 23 states, resulting in the low error estimate of .1865. Each state in the optimal sequence had an average of .033 possible alternatives. This resulted in a confidence estimate of 96.7 percent. The lower face, as well, was trained 105 different times. The average HMM among these contained 24 independent states.





The error produced by the average HMM was .3289. Each state had, in the optimal sequence, an average of .052 possible alternatives, producing a confidence estimate of 94.2 percent.

The entire face was, once again, trained 105 different times. The smallest average error of .3618 was obtained with an HMM that was composed of 23 states and five dimensions. Each state possessed an average of .042 alternatives from the optimum path of probabilities, allowing a confidence estimate of 95.8 percent. Statistical analysis of variance in the holistic model was  $p = 10^{-5}$ .

## Discussion

The compactness of any model is based on the amount of data needed to run it. In the case of the FACS, the action units needed, the extent to which each will be used, the direction of the move, and the duration of the movement must be specified for each viseme. This large amount of information required reveals that the FACS-

derived model is not compact. In contrast, the HMM model developed from the algorithm requires only two sets of time-independent information to simulate a given sequence: initial probabilities and transitional probabilities. Given these, it is possible for the HMM to perform a fast and efficient sequencing of facial processes. While the HMM uses an average of 23 states in contrast to the FACS-derived model's use of seven action units, this does not impact the complexity of the FACS vs. the HMM as the units are not equivalent. Each state of the algorithm-derived HMM compresses two or more fully defined action units into simpler terms of coordinate movement and velocity, allowing for a compact representation of the same material in flexible terms dependent only on the vocal and facial peculiarities of the subject.

Compact models of a process should imitate the internal structure of the facial processes. This is revealed primarily through three major components of muscles: time, subtlety, and interdependencies. The FACS-derived model stresses static expressions by ignoring the time component of facial expression. This implies that facial expression is solely dependent on the position of facial features when, in fact, motion plays a significant role in distinguishing between emotions, as stated before. Sequencing and timing of expressions are especially critical to the recognition of emotions. These facts are acknowledged by the HMM, which includes velocity in its assessment of states. By doing so, it bypasses the inability of the FACS-derived model to deal with the timing of the separate types of muscles found in the face: linear, sheet, and sphincter. This ability of the HMM to specify the velocity as well as the position of landmark features promotes natural movement as velocities are assigned to the appropriate regions of the face within the proper context.

In addition to neglecting the dynamics of muscle motion, the FACS neglects to address the subtlety of muscle motion. The FACS stresses large, local, and isolated spatial patterns. It is unable to describe fine eye or lip motions, concentrating primarily in gross motion. In addition, it is unable to predict the physiological effect that muscles have on each other. The need for fine detail is emphasized by the lip movement shown by the 8-state HMM, which demonstrates the fine detail needed in differentiating phonemes. The HMM predicts the face holistically, defining each state for the entire face—not for a portion of it—so logical side-effects are encoded along with the intended movement. In addition to simplifying simulation, the trained HMM unobtrusively

absorbs mechanisms, hidden or known, that regularly correspond to given movements.

The inability of FACS-dependent models to address holistic movements is displayed prominently in the discussion of coarticulation. Since coarticulation, as previously mentioned, involves the modification of visemes based on its context, it poses a uniquely difficult problem for FACS-derived models. The need for context has been well documented for its ability to assure natural movement, but the FACS is unable to provide for context due to its emphasis on static poses.<sup>26,27</sup>

The HMM, however, is able to meet this need. By definition, transitional probabilities preserve context within HMMs by determining the circumstances under which a designated state should appear. As the decision of a succeeding state is entirely independent of the preceding sequence, context is the only deciding factor that affects the choice of the successor. This property of HMM allows it to effectively mimic the mechanics of facial processes. The entropy rates of the HMMs developed indicate that context is used both before and after designated sequences in a time frame similar to that found in facial and visual coarticulation. The FACS-derived model copes by assigning elaborate algorithms and dominance rankings to deal with coarticulation (MIKETALK) or by simplifying the viseme-to-phoneme relationship to a one-to-one arrangement. Those methods, however, cannot overcome this internal defect inherent in the FACS.

Qualitatively, the velocity trajectories developed for each state were smooth and similar to natural facial movement. Resulting transitions from state-to-state were geometrically and temporally well-behaved.

A compact model of a process that contains mechanisms intrinsic to an original process must be able to extrapolate sets consistent with output from that process. The FACS-derived model does not contain many of the mechanisms found in

the face, and it has a correspondingly high error estimate of .6881. The HMM, however, contains some of the mechanisms found in the face, and has an overall low error. The error estimate for the entire face, .3618, was the sum of the error estimate of the lower face (.3289) and the upper face (.1865). The confidence estimate for states was 95.9 percent for the entire face, which was the average of the lower face (94.2%) and upper face (96.7%) confidence estimates. On the whole, these low estimates confirm that the common mechanism found in the face and the model produce similar data sets, as would be expected.

## Conclusions and Open Questions

Previously, Kolmogorov complexity supported the claim that a compact model is capable of reproducing the internal structure of a process. This was validated by the study, which verified the quantitatively superior capability of the HMM model in accuracy, compactness, and pattern to a model using standard encoding sequences (the FACS). The developed HMM allows for a greater understanding of facial processes and a viable process for quantitatively simulating the face. Further experiments extending the range of subjects and expressions may be useful.

The HMM model has numerous applications. By determining the structure of the face, it may allow an understanding of the neuromuscular structure and additional hidden mechanisms of the face. This may prove useful in cases of facial paralysis, such as that found in strokes and in Moebius Syndrome. In addition, it may provide further information as to the correlation between vocal information and facial expressions. Its unique probability distributions specific to training data and facial peculiarities will also allow further applications in data compression. ■