Learning expressive human-like head motion sequences from speech

Carlos Busso[1], Zhigang Deng[2], Ulrich Neumann[1], and Shrikanth Narayanan[1]

[1] Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA
busso@usc.edu, shri@sipi.usc.edu, uneumann@graphics.usc.edu
[2] Department of Computer Science, University of Houston, Houston, TX 77204, USA
zdeng@cs.uh.edu

1 Introduction

With the development of new trends in human-machine interfaces, animated feature films and video games, better avatars and virtual agents are required that more accurately mimic how humans communicate and interact. Gestures and speech are jointly used to express intended messages. The tone and energy of the speech, facial expression, rigid head motion and hand motion combine in a non-trivial manner as they unfold in natural human interaction. Given that the use of large motion capture datasets is expensive and can only be applied in planned scenarios, new automatic approaches are required to synthesize realistic animation that capture and resemble the complex relationship between these communicative channels. One useful and practical approach is the use of acoustic features to generate gestures, exploiting the link between gestures and speech.

Since the shape of the lips is determined by the underlying articulation, acoustic features have been used to generate visual *visemes* that match the spoken sentences [4, 5, 12, 17]. Likewise, acoustic features have been used to synthesize facial expressions [11, 30], exploiting the fact that the same muscles used for articulation also affect the shape of the face [44, 46]. One important gesture that has received less attention than other aspects in facial animations is rigid head motion.

Head motion is important not only to acknowledge active listening or replace verbal information (e.g. "nod"), but also for many aspect of human

[1] This is a preprint of a book chapter to appear at "Data-Driven 3D Facial Animation", Zhigang Deng and Ulrich Neumann, Springer-Verlag Press, 2007, (in print). http://www.springer.com/west/home?SGWID=4-102-22-173735422-0&changeHeader=true&SHORTCUT=www.springer.com/978-1-84628-906-4

communication (for details see [26]). Graf *et al.* suggested that rigid head motion is used to segment the linguistic units of spoken content, since the timing between the prosodic structure and head motion are consistent [23]. Head motion also improves acoustic perception, as noted by Munhall *et al.* [36]. They also suggested that head motion helps to distinguish between interrogative and declarative statements. Hill and Johnston show that head motion is used to recognize speaker identity [27]. Moreover, Jefferies *et al.* suggest that head motion influences the perception of the personality of the animated character. Similarly, our previous work indicates that head motion affects the emotional perception of facial animations [6].

Given the importance of head motion in human-human interaction, this non-verbal channel needs to be properly modeled for realistic facial animation. Kuratate *et al.* have estimated the correlation levels between prosodic and head motion features [32]. Based on the high correlation levels achieved ($r = 0.8$), they concluded that the production of speech and head motion are internally linked. Even though head motion patterns depend on many other factors such as the underlying semantic content and the personality of the subjects, these results suggest that speech can be used to generate head motion sequences.

In this chapter, the relationship between rigid head motion and prosodic speech is analyzed, in terms of emotional categories (neutral state, sadness, happiness and anger). The results show that head motion and prosodic speech are strongly connected. However, the relationship varies from emotion to emotion, suggesting that emotional models need to be built to generate realistic head motion sequences. Based on this study, a novel approach to synthesize head motion sequences from prosodic speech is presented. In this framework, head poses are quantized in a finite number of clusters or codebooks. For each of these codebooks, a *Hidden Markov Model* (HMM) is built, taking prosodic features as observations. In the synthesis step, the acoustic features of the test speech are entered in the HMMs and the most likely head motion sequences are generated. Smoothing techniques based on first order Markov models followed by spherical cubic interpolation are used to ensure continuous head motion sequences. To include emotional patterns in the generated sequences, different sets of HMMs are built for each emotional category. Evaluations of this framework reveal that the generated sequences follow the temporal dynamics of speech well. Moreover, the generated sequences were judged by human raters at the same level of naturalness as the captured head motion sequences. Previous versions of this framework were published in [6, 7].

This chapter is organized as follows: Section 2 presents previous work on head motion synthesis. It also motivates the importance of modeling emotion for engaging animated characters. Section 3 describes the audio-visual database and the procedure used to extract the audio-visual features. In Section 4, the relationship between head motion and prosodic features is analyzed in terms of emotional categories. Section 5 describes the framework used to synthesize head motion sequences. Section 6 presents the objective and sub-

jective evaluations of this approach. Finally, Section 7 gives the concluding remarks and our future research directions.

2 Related work

2.1 Head motion synthesis

Different approaches have been used to synthesize head motion sequences, given the relationship between head motion and the verbal message. For instance, plain text enriched with manual annotations of discourse functions were used to synthesize well-known head motion gestures such as head "tilt" and "nod". De Carlo *et al.* present a coding-based platform for real-time facial animation that supports head motion rotation and translation [14]. The movements of the head were driven by manual annotations of specific head motion gestures co-occurring with prominent words in the text. Pelachaud *et al.* propose a rule-based system to synthesize head movement from text responding to discourse function (e.g. conversational signal and punctuators) [37]. If the emotion of the animation were specified, the velocity and the global pose of the head motion were modified according to pre-defined rules (e.g. for sadness, the global pose was set with downward direction). Similar rule-based systems are presented in [10, 43].

Instead of using text, other approaches have been proposed to exploit the prosodic structure of the acoustic signal. The prominence in speech prosody is closely related to head motion [10, 24, 32]. Therefore, it can be used to estimate head motion sequences. Albrecht *et al.* propose the generation of head movements based on the pitch contour [1]. If the difference between two maxima of the pitch exceeded a threshold, the head was raised. If the difference between two minima exceeded a given threshold, the head was lowered. The amplitude of the upward and downward movements was proportional to the magnitude of the differences. Random movements in the horizontal and vertical axes were added to prevent repetitive movements.

Graf *et al.* analyze the relationship between head motion and the prosodic structure in the speech [23]. They define few primitives to describe head motion (e.g. "nod") that were consistently observed across speakers. The co-occurrence between these primitives and prosodic events, which were labeled using the *Tones and Break Indices* (ToBI) scheme, are used to estimate the conditional probability of major head movement given pitch accents. A similar approach is presented by Sargin *et al.* , in which specific head motion sequences for "nod" and 'tilt" are generated when pitch accent is detected [40]. Unfortunately, these approaches only generate limited head motion gestures, which do not reflect the wide range of head motion patterns displayed during human-human interaction.

Chuang and Bregler present a data-driven approach to synthesize head motion sequences [11]. In this approach, the head motion and pitch contour

corresponding to segments in the training data were recorded. The prosodic structure of each new speech signal is compared with the ones in the database with similar emotional content. After selecting the top M matches for each segment in that sentence, a dynamic programming algorithm was used to find the optimum path of the head motion sequences. The cost function was designed to achieve smooth transitions between segments. Deng *et al.* developed a similar head motion synthesis technique [16]. In addition to searching the M best matches between the novel speech material and the ones in the database, they proposed the inclusion of optional key framing controls. With this extension, the designers were able to incorporate specific rigid head gestures in the animation, such as "head nod". Then, a dynamic programming algorithm maximized the optimum path between the head motions segments, constrained by the specified key head poses. One advantage of these two studies is that the head motion sequences are not restricted to a few prototype rigid head gestures.

In the proposed framework, the focus is on modeling the temporal relationship between head motion and prosodic features. HMMs are used to estimate discrete representation of head poses from prosodic features. As shown in Section 6, the resulting head motion sequences preserve the temporal relation between head motion and speech. More importantly, in the context of facial animation, human evaluators perceived them as natural as the captured head motion sequences.

2.2 Emotion in facial animation

For engaging talking avatars, special attention needs to be given to include emotional capability in the virtual characters. Importantly, Picard has underscored that emotions play a crucial rule in rational decision making, in perception and in human interaction [38]. In fact, Gratch and Marsella have proposed the use of emotions as a crucial component in the decision-making model of human-like characters [25]. Therefore, applications such as virtual teachers, animated films and new human-machine interfaces can be significantly improved by designing control mechanisms to animate the character to properly convey and react according to the desired emotion. Human beings are especially good not only at inferring the affective state of other people, even when emotional clues are subtly expressed, but also in recognizing non-genuine gestures, which challenges the designs of these control systems.

The production mechanisms of gestures and speech are internally linked in the brain. Cassell *et al.* mention that these mechanisms are not only strongly connected, but also systematically synchronized in different scales (phonemes-words-phases-sentences) [10]. They suggest that hand gestures, facial expressions, head motion, and eye gaze occur at the same time as speech, and convey information similar to that in the acoustic signal. Similar observations are mentioned by Kettebekov *et al.* [31]. They studied *deictic* hand gestures (e.g. pointing) and speech prosody in the context of gesture recognition. They

concluded that there is a multimodal coarticulation of gestures and speech, which are loosely coupled.

From an emotional expression point of view in communication, it has been observed that human beings jointly modify gestures and speech to express emotions. Communicative channels such as facial expressions [21, 22], head motion [6, 37], pitch [13, 41] and short time spectral envelope [47] all present specific patterns under emotional states. Therefore, a more complete human-computer interaction system should include details of the emotional modulation of gestures and speech.

In sum, all these findings suggest that the control system to animate virtual human-like characters needs to be closely related and synchronized with the information provided by the acoustic signal. In addition, these control systems need to model the emotional content that the animated character is supposed to convey. This chapter proposes the use of emotion-dependent models driven by prosodic features to synthesize realistic head motion sequences.

3 Audio-visual database

The audio-visual database was recorded from an actress who was asked to read a semantically-neutral, custom-made, phoneme-balanced corpus four times, expressing different emotions: neutral state, sadness, happiness and anger, at each reading. Facial markers were attached to her face according to the layout illustrated in Fig. 1. The markers were tracked with a VICON motion capture system with three cameras at a sampling rate of 120 frames per second (right of Fig. 1). Her speech was simultaneously recorded with a close talking SHURE microphone at 48Khz. In total, 640 sentences were used in this work. Notice that the actress was instructed to act naturally without any specific instructions about how to move her head.

After the data were collected, the markers' positions were translated to make the lower nose marker the center of the coordinate system. After that, the three degrees of head rotation were estimated using a technique based on *Singular Value Decomposition* (SVD) [2]. In this technique, a reference frame was selected from a neutral pose. The 3D position of the markers were arranged as a 102×3 matrix, referred here on as M_{ref}. Each of its rows contains the x, y and z location of the markers. Then, for frame t, a matrix M_t is created following the same order as in M_{ref}. Then the SVD, UDV^T, of the matrix $M_t \cdot M_{ref}$ is calculated. The product VU^T gave the rotation matrix for frame t.

$$M_{ref}^T \cdot M_t = UDV^T \tag{1}$$

$$R_t = \ VU^T \tag{2}$$

Finally, head motion is modeled as the 3D Euler angles, x_t, corresponding to head rotation, which are derived from R_t (Fig. 2). Notice that head motion

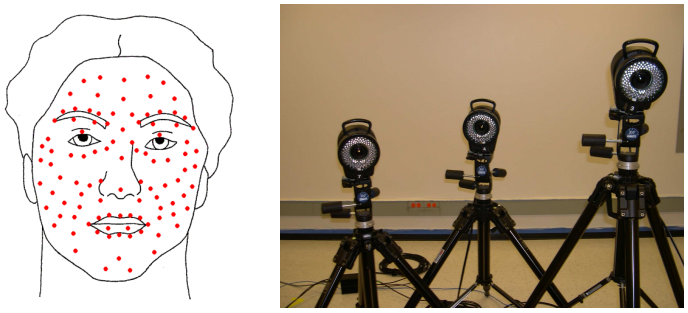


Fig. 1. Audio-visual database. The left figure shows the layour of the 102 facial markers, and the right figure shows the motion capture system.

is usually parameterized with 6 *degrees of freedom* (DOF), corresponding to rotation (3 DOF) and translation (3 DOF) [16, 45]. As discussed in Section 5, the proposed framework requires discrete head poses, which are estimated using vector quantization. For a constant quantization error, the number of clusters significantly increases if 6 DOF are considered instead of 3 DOF. Fortunately, from a practical point of view, most applications require close-view of the face, in which translation effects are less important than rotation effects.

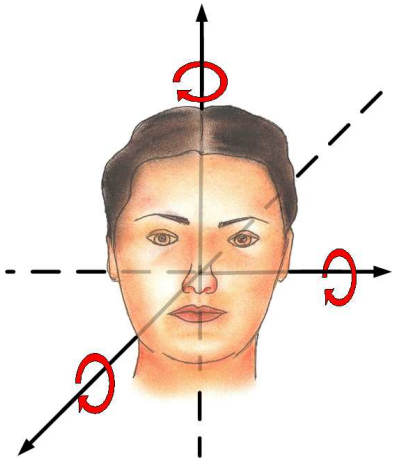


Fig. 2. Head motion parameterization.

The acoustic features were extracted using the Praat speech processing software [3]. The analysis window was set to 25 milliseconds with an overlap of 8.3 milliseconds, producing 60 frames per second. The pitch (F0) and the RMS energy were estimated. The pitch was smoothed to remove any spurious spikes, and interpolated to avoid zeros in the unvoiced region of the speech, using the corresponding options provided by the Praat software. The first and second derivatives of these features were also considered, since they provided useful temporal information. In sum, a 6D feature vector was used. Notice that prosodic features predominantly describe the source of the speech rather than the vocal tract. Therefore, this head motion framework is independent of the specific lexical content of the sentence, reducing the size of the database needed to train the models.

4 Analysis of head motion during expressive speech

In this section, a brief analysis of the relationship between head motion and prosodic features in terms of emotional categories is studied. For this purpose, the database was split into emotional categories, and different statistical measures were computed. The main goal of this study is to quantify differences in the head motion patterns displayed under expressive utterances.

Table 1. Statistics of head rotation [6].

	Neu	Sad	Hap	Ang
		Canonical correlation Analysis		
	0.74	0.74	0.71	0.69
		Motion Coefficient [°]		
α	3.32	4.76	6.41	5.56
β	0.88	3.23	2.60	3.67
γ	0.81	2.20	2.32	2.69
		Range [°]		
α	9.54	13.71	17.74	16.05
β	2.31	8.29	6.14	9.06
γ	2.27	6.52	6.67	8.21
		Velocity Magnitude [°/sample]		
Mean	.08	0.11	0.15	0.18
Std	.07	0.10	0.13	0.15

To measure the relationship between head motion and prosodic features, *Canonical Correlation Analysis* (CCA) [28] was applied to the data. CCA provides a scale-invariant optimal linear framework to measure the correlation between two streams of data with equal or different dimensionality. In this method, the feature vectors are projected into a common space in which

Pearson's correlation can be measured. Table 1 shows the average first order canonical correlation between head motion and speech. The results show correlation levels higher than $r = 0.69$ across emotional categories. This result agrees with the observation made in [32], about the close relation between head motion and prosodic features. Notice, however, that the correlation levels vary from emotion to emotion [8]. A one-way *analysis of variance* (ANOVA) evaluation indicates that there are significant differences between emotional categories (F[3,640], $p = 0.0013$). In fact, multiple comparison tests reveal that the average CCA of neutral head motion sequences is different from the average CCA of sadness ($p = 0.001$) and anger ($p = 0.001$).

To measure the level of head movement activity, a *motion coefficient*, Ψ, is defined as the standard deviation of the sentence-level mean-removed signal,

$$\Psi = \sqrt{\frac{1}{N \cdot T} \sum_{u=1}^{N} \sum_{t=1}^{T} (x_t^u - \overline{\mu^u})^2} \tag{3}$$

where T is the number of frames, N is the number of utterances, and $\overline{\mu^u}$ is the mean of the sentence u. The average results of this *motion coefficient* when applied to head motion features are presented in Table 1. The results indicate that head motion activity during emotional utterances is significantly higher than in neutral utterances. Happiness and anger present the highest levels of head motion activity. As an aside, it is interesting to notice that similar trends were also observed in the articulatory domain for tongue and jaw movement [33].

Table 1 also gives the average range and velocity of expressive head motion patterns. These results indicate that during emotional utterances the head moves over a wider range than in the neutral case. Likewise, for happiness and anger the head motion velocity increase more than 90%, compared to the neutral case. These results, which agree with previous work [37], indicate that the temporal dynamics of head motion during neutral speech presents important differences compared to the patterns displayed during emotional speech.

A discriminant analysis was applied to the data, to infer how distinct the head motion patterns are under different emotional categories. The mean, standard deviation, range, maximum and minimum of the head motion features computed at the sentence-level were used as features. Fisher classification was implemented with leave-one-out cross validation method. Table 2 shows the results. On average, the recognition rate was 65.5% using only head motion features. Notice that the emotional class with the lowest performance (anger) is correctly classified with accuracy higher than 50% (chance is 25%). These results suggest that there are distinguishable emotional characteristics in rigid head motion. Also, the high recognition rate of neutral state implies that global patterns of head motion in normal speech are completely different from the patterns displayed under an emotional state.

Table 2. Emotional discriminant analysis of head rotation [6].

	Neu	Sad	Hap	Ang
Neu	0.92	0.02	0.04	0.02
Sad	0.15	0.61	0.11	0.13
Hap	0.14	0.09	0.59	0.18
Ang	0.14	0.11	0.25	0.50

Previous work has shown than prosodic features are also affected by emotional modulation [13, 41]. As a result, it is not surprising that the relationship between head motion and prosodic features is emotion-dependent.

These results agree with our previous work, which indicates that head motion is one of the facial gestures that is less constrained by articulatory processes [9]. As a result, it can be used with less restriction to express other non-linguistic messages, such as emotions. Therefore, emotion-dependent head motion models are needed for human-like expressive facial animation. Further details on the analysis can be found in [6, 8, 9].

5 Head motion framework

As discussed in the previous section, head motion and prosodic features are closely related across time. *Hidden Markov Model* (HMM) is a statistical time-series framework that has been used to model similar data. Accordingly, we propose to generate head motion sequences using HMMs. Instead of estimating a mapping function [23], designing rules according to specific comunicative functions [10, 37, 43], or finding similar samples in the training data [11, 16], we model the problem as classification of discrete representations of head motion using acoustic prosody as feature. That is to say, the relationship between head motion and prosodic features is directly learned from data, without specifying the high-level functions in the speech. We will discuss this point further in Section 7.

Since an HMM will be built for each head pose, a discrete representation of head motion is needed. This representation is obtained by using the *Linde-Buzo-Gray Vector Quantization* (LBG-VQ) technique [34]. The 3D space spanned by the head motion features is split in K Voronoi cells. For cell V_i with $i \in \{1, \ldots, K\}$, the mean, U_i, and covariance matrix Σ_i of the points inside the cluster are estimated (Fig. 3). In the quantization step, the continuous Euler angles of each frame are approximated with the closest vector code in the codebook.

For each of the head pose cluster (V_i), an HMM is built to generate the most likely head motion sequence, given the observation O, which corresponds to the prosodic features. Therefore, the number of models that will be built is given by the number of clusters (K) used to represent the head poses. The HTK toolkit is used to build these HMMs [48].

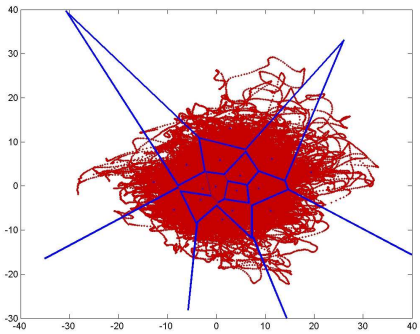


Fig. 3. 2D projection of the Voronoi regions using 16-size vector quantization.

To guarantee a continuous head motion sequence without breaks, two smoothing constraints are imposed. The first smoothing technique takes place in the decoding step of the HMMs. In this approach, transition between head motion cluster are constrained according to their appearance in the training data. The second smoothing constraint is imposed as a post-processing step, by using spherical cubic interpolation. Further details of these smoothing techniques are given in Sections 5.1 and 5.2, respectively.

As discussed in Section 4, the relationship between head motion and prosodic features depends on the emotional content of the utterance. If human-like facial animations are required, the specific emotional patterns of the gestures, in this case head motion, need to be appropriately designed. In this proposed approach, the relationship between head motion and prosodic features is learned in terms of emotional categories. The data is split according to the emotional labels, and emotion-dependent HMMs are separately built. Therefore, the specific emotional patterns are directly included in the models.

5.1 Learning head motions

To synthesize human-like head motion, our technique searches for the sequences of discrete head poses that maximize the posterior probability of the cluster models $V = (V_{i_1}^t, V_{i_2}^{t+1}, \ldots)$, given the observations $O = (o^t, o^{t+1}, \ldots)$.

$$\arg \max_{i_1, i_2, \ldots} P(V_{i_1}^t, V_{i_2}^{t+1}, \ldots | O) \tag{4}$$

Instead of directly modeling this posterior probability, the problem is solved by training the prior probability, $P(O)$, and the likelihood, $P(O|V_{i_1}^t, V_{i_2}^{t+1}, \ldots)$, by making use of Bayes' rule:

$$P(V|O) = \frac{P(O|V) \cdot P(V)}{P(O)} \tag{5}$$

The likelihood distribution, $P(O|V)$, models how well the head motion models fit the data. Here, it is modeled as a first order Markov process with S states. Therefore, the probability description at time t includes only the current and the previous states, which significantly simplifies the problem. In each of the states, the distribution of the observations are modeled with a *Mixture of M Gaussians*. As noted in [6, 7], the mapping between head motion and prosodic features is *many-to-many*. By using Mixture of Gaussians to model the distribution of the observations, this ambiguous relationship is included in the models. Under this formulation, building the likelihood distribution of head motion sequences is reduced to learning the parameters of standard HMMs. For this training problem, well-known techniques such as forward-backward and Baum-Welch re-estimation algorithms are used. For more information about HMM, the readers are referred to [39, 48].

The prior probability, $P(V)$, in Equation 5 plays an important role in this framework. It models the transition probability between head motion clusters based only on prior information. Here, $P(V)$ is used as a first smoothing technique to guarantee valid transitions between the discrete head poses. Similar to bi-gram models used for language models [48], this prior probability is modeled as a first-order state machine. The transitions between clusters are learned by counting their relative frequency in the training data. In the decoding step of the HMMs, this prior information is used to reward or penalize transitions that are frequently or seldom observed in the database, respectively. According to the analysis presented in Section 4, head motion dynamics are also affected by the underlying emotion of the subject. Therefore, this prior probability is separately learned from each emotional category.

P(O) does not depend on the head motion models and is a constant in Equation 4. Therefore it can be ignored in this framework.

Notice that in the training procedure the segmentation of the acoustic signal is obtained from the vector quantization of the head motion space. Therefore, the HMMs were initialized with this known segmentation, avoiding the use of forced alignment, as it is usually done in automatic speech recognition to align phonemes with the speech features.

5.2 Head motion synthesis

Figure 4 describes the proposed framework to generate human-like head motion sequences. After the HMMs are trained, the prosodic features described in Section 3 are extracted from the acoustic signal of the test database. This feature vector is used as an observation of the HMMs, which generates the most likely sequences of head poses codebooks, $\widehat{V} = (\widehat{V}_{i_1}^{t}, \widehat{V}_{i_2}^{t+1} \ldots)$, according to Equation 5. After the sequence $\widehat{V}$ is generated, the means of the clusters are used to form a 3D sequence, $\widehat{Y} = (U_{i_1}^{t}, U_{i_2}^{t+1} \ldots)$, which is the first approximation of the head motion.

The next step in this approach is to blur the sequence $\widehat{Y}$ with additive colored noise (Equation 6). The purpose of this step is to compensate for

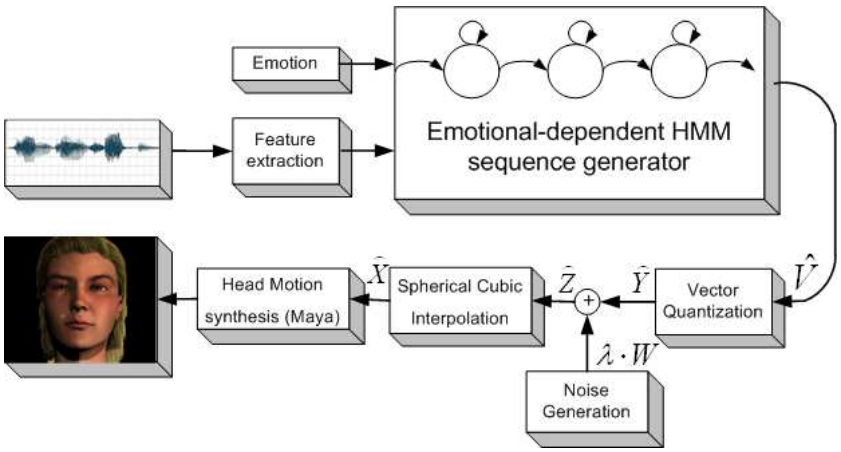


Fig. 4. Emotion-dependent head motion synthesis framework.

the quantization error yielded during vector quantization. Hence, the noise is added such that the covariance of the noise matches the covariance matrix associated with the codebooks (Σ). Therefore, the power of the noise is distributed in proportion of the quantization error. The parameter λ is included in Equation 6 to attenuate, if desired, the noise level used to blur the sequence $\widehat{Y}$ (e.g. $\lambda = 0.7$). Notice that this is an optional step that can be ignored by setting λ equal to one, if no attenuation is desired, or to zero if no noise is desired. The solid blue line in Fig. 5 shows an example of the noisy version of the head motion sequences, $\widehat{Z}$.

$$\widehat{Z}_i^t = \widehat{Y}_i^t + \lambda \cdot W(\Sigma_i) \tag{6}$$

As can be observed from Fig. 5, the noisy version of the head motion sequence ($\widehat{Z}$) presents a break in the cluster transitions. This problem is observed even when the number of codebooks or the noise level is increased (K). To avoid these discontinuities, a second smoothing technique is applied to the sequence. If a standard cubic interpolation is separately applied to each of the Euler's angles, it is well known that the resulting sequence may present jerky movements and undesired effects such as *Gimbl lock* [42]. Instead, the proposed smoothing technique is based on spherical cubic interpolation [20]. With this technique, the 3D Euler angles are jointly interpolated in the unit sphere by using quaternion representation, avoiding the artifact mentioned before.

In the interpolation step, the sequence $\widehat{Z}$ is downsampled to 6 points per second to obtain equidistant frames. These frames are referred, from here on, as key-points and are marked in Figure 5 as a black circle. These 3D Euler angle points are then transformed into the quaternion representation [20]. Spherical cubic interpolation (squad) is then applied over these quaternion points.

The squad function builds upon the spherical linear interpolation (slerp). The functions slerp and squad are defined by equations 7 and 8, respectively:

$$\text{slerp}(q_1, q_2, \mu) = \frac{\sin(1-\mu)\theta}{\sin\theta} q_1 + \frac{\sin\mu\theta}{\sin\theta} q_2 \quad (7)$$

$$\text{squad}(q_1, q_2, q_3, q_4, \mu) = \text{slerp}(\text{slerp}(q_1, q_4, \mu),\ \text{slerp}(q_2, q_3, \mu), 2\mu(1-\mu)) \quad (8)$$

where q_i (with $i \in \{1, \ldots, 4\}$) are quaternions, $cos\theta = q_1 \cdot q_2$ and μ is a parameter that ranges between 0 and 1 and determines the frame position of the interpolated quaternions. With these equations, the head motion sequence is interpolated in the unit sphere by varying the parameter μ to recover the original sample rate (120 frames per second). The last step in this smoothing technique is to transform the interpolated quaternions into an Euler angle representation.

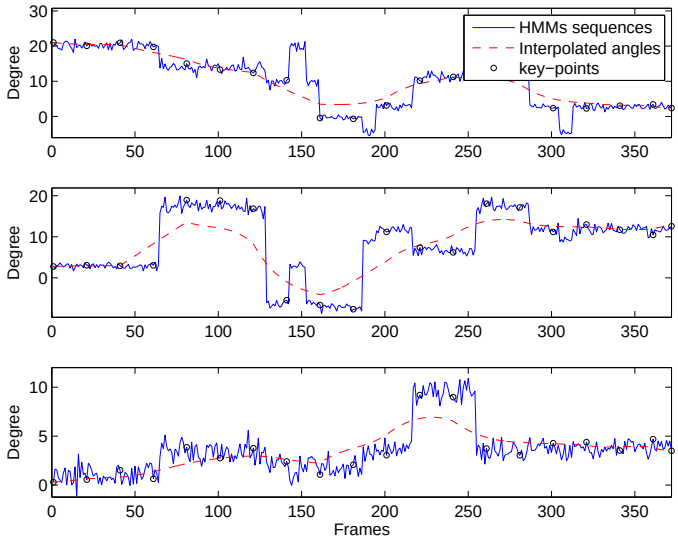


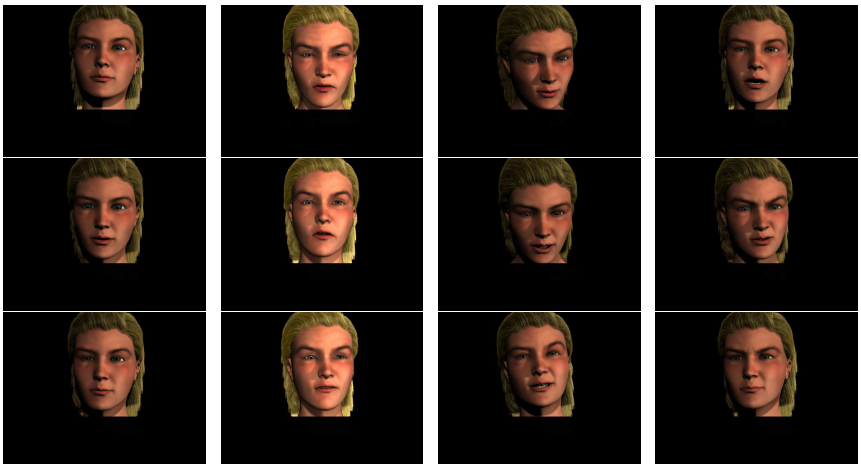
Fig. 5. Example of a synthesized head motion sequence. The solid blue line represents the 3D noisy signal $\hat{Z}$ from equation 6. The circles are the key points used for spherical cubic interpolation. The dashed red line is the smoothed head motion sequences used in the animation ($\hat{X}$).

Notice that the noise is added before the spherical cubit interpolation technique is applied. Therefore, the resulting head motion sequence, referred by $\widehat{X}$, is a continuous and smooth 3D signal without the jerky behavior of noise. An example of this sequence is illustrated in Fig. 5 as dashed red line.

The final step in this framework is to include the head motion sequence $\widehat{X}$ in the facial animation. Here, a blend shape face model composed of 46

blend shapes is used, which is modeled and rendered using Maya [35]. The head motion sequence $\widehat{X}$ is directly applied to the angle control parameters of the face model. For realistic human-like expressive animation, other facial components such as expressive visual speech and eye motion are also modeled and synthesized. The details of those approaches can be found in [15, 17–19].

Figure 6 shows frames of the synthesized sequences for each of the four emotional categories considered here.



(a) Neutral state (b) Sadness (c) happiness (d) anger

Fig. 6. Frames from synthesized sequences. Each column corresponds to a specific emotional category.

5.3 Parameter configuration of the models

An important parameter in this framework is the HMM topology, which is defined by the number and interconnection of the states. The most common topologies are the *left-to-right topology* (LR), in which only transitions in forward direction between adjacent states are allowed, and the ergodic topology (EG), in which the states are fully connected. In the LR topology, fewer parameters are required. Therefore, less data is needed to train its parameters. The EG topology is less restricted, so it can learn the state transitions from the database. However, more parameters are needed to learn the models, which increase the requirement on the size of training data.

In this particular problem, it is not clear which HMM topology provides the best description of the head motion dynamics. In our previous work, different generic HMM configurations were evaluated. By generic models, we mean emotion-independent HMMs that were learned without considering the emotional category of the sentences (the entire database was used for training).

The Left-to-Right HMM, with 3 states (S) and 2 mixtures (M) achieved the best result. Notice that if the database were big enough, an ergodic topology with more states and mixtures could perhaps give better results.

When emotional models are used instead of generic models, the training data is even smaller, since the emotion-dependent models are separately trained. Therefore, the HMMs used in the experiments were implemented using a LR topology with 2 states ($S = 2$) and 2 mixtures ($M = 2$).

Another important parameter of this model is the number of clusters (K) used to create the discrete representation of head motion, which is directly related with the number of HMMs that is to be built. If K increases, the quantization error of the discrete representation of head poses decreases. However, the discrimination between models will significantly decrease and more training data will be needed. Therefore, there is a tradeoff between the quantization error and the inter-cluster discrimination. As shown in [6, 7], a 16 word codebook was adequate to synthesize realistic facial animation. In the experiments reported here, K was also set to 16.

The audio-visual database mentioned in Section 3 was randomly split in training (80%) and testing data (20%).

6 Head motion evaluation

The head motion sequence framework presented here was objectively and subjectively evaluated. This section presents the main results.

6.1 Objective evaluation

The first order canonical correlation between the original and synthesized head motion sequences was computed to analyze whether this framework is able to capture the temporal relationship between head motion and prosodic features. The average results are presented, separately for each emotional category, in Table 3. The results show that the sequences generated with prosodic features are highly correlated with the original captured sequences. This suggests that this framework appropriately models the temporal behavior of head motion sequences. Notice that the first order correlation between head motion and prosodic features was about $r \approx 0.72$ (Table 1). Interestingly, the first order canonical correlation between the original and synthesized head motion sequences was over $r > 0.85$. Even though prosodic features do not provide all the information needed to synthesize head motion sequences, this result indicates that the performance of the proposed system is notably high.

6.2 Subjective evaluation

For subjective assessments of this framework, 17 human subjects were asked to rate the naturalness of videos with facial animation rendered with the

Table 3. Canonical Correlation Analysis between original and synthesized head motion sequences [6].

	Neutral	Sadness	Happiness	Anger
Mean	0.86	0.88	0.89	0.91
Std	0.12	0.11	0.08	0.08

synthesized and original head motion sequences. Likewise, they were also asked to rate the naturalness of the animation without head motion (rigid head), to study how important head motion is for realistic facial animation. In total, one video per each emotional category was presented to the evaluators, resulting in 12 videos: 4 emotions $\times$ 3 modalities (synthesized, original and rigid). Although the facial animations included other facial gestures such as lips and eyes, the only gesture that was modified was head motion.

The animations were presented in random order. The naturalness of the animation was rated using a five-point scale. The extremes were called *robot-like* (value 1), and *human-like* (value 5). The evaluators received instructions to rate their overall impression of the animation and not individual aspects such as head movements, or voice quality. The subjects were not made aware that head motion was the component of the facial animation that was under assessment.

Table 4. Naturalness assessment of rigid head motion sequences [1-*robot-like*, 5-*human-like*] [6].

Head Motion Data	Neutral		Sadness		Happiness		Anger	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Original	3.76	0.90	3.76	0.83	3.71	0.99	3.00	1.00
Synthesized	4.00	0.79	3.12	1.17	3.82	1.13	3.71	1.05
Fixed Head	3.00	1.06	2.76	1.25	3.35	0.93	3.29	1.45

Table 4 presents the results for the subjective assessment. With the exception of sadness, the synthesized sequences were judged to be more natural than the animation with the original head motion sequences. This result indicates that the head motion synthesis approach presented here was able to generate realistic human-like head motion sequences. One aspect that significantly improves the naturalness perception of the facial animation is the synchronism between the prosodic structure of the speech and the head motion. Prominence in the speech was systematically accompanied with head motion gestures, which indicates that this framework was able to model the non-trivial relationship between head motion and speech.

Table 4 also shows how the listeners assessed the naturalness of the facial animation without head motion. These results show that the naturalness

perception significantly decreases when head motion is not included in the facial animation. This implies that head motion is an important component for human-like facial animations that needs to be appropriately modeled for engaging animated characters.

The inter-evaluator average variance in the scores rated by human subjects was 0.97. This result indicates that the concept of naturalness of the animation is perceived slightly differently between the evaluators. However, since we are interested in the mean differences of each group, this variability does not bias these results.

7 Discussion and Conclusions

Head motion is an important component in interpersonal human interactions. Therefore, this gesture needs to be properly modeled and included in realistic human-like facial animations. The subjective evaluations presented in this chapter support this observation, since the naturalness perception significantly decreases when the animations were rendered without head motion.

As analyzed in this chapter, the head motion patterns displayed under expressive utterances vary across emotional categories. For instance, the range and velocity of the head present higher values for happiness, anger and sadness, compared with the values for neutral utterances. In fact, the results reveal that head motion can be used to discriminate between emotional categories. Since prosodic features are also affected by the affective state of the subject, the relationship between head motion and prosodic features is emotion-dependent. This observation is supported by the canonical correlation analysis, which indicates that the correlation levels between head motion and prosodic features are significantly different across emotional categories. Furthermore, our previous work indicates that head motion influences the emotional perception of facial animations [6]. As results, emotion-depend head motion models need to be designed for human-like facial animation.

Based on previous observations, a novel data-driven framework to synthesize head motion sequences based on prosodic features was presented. In this technique, discrete representations of head poses, estimated with vector quantization, were modeled with HMMs, which took prosodic features as inputs. A set of HMMs was separately trained for each emotional category, building emotion-dependent head motion models. The subjective and objective evaluations indicate that this framework successfully modeled the temporal relationship between head motion and speech. Furthermore, the facial animations with the synthesized head motion sequences were perceived having the same level of naturalness as the animations with the motion captured head motion sequence.

In this work, head motion is only modeled with 3 DOF corresponding to head rotation. However, the human neck allows the head not only to rotate,

but also translate, especially back and forward. An interesting question is how to include in this framework this extra 3 DOF of head translation.

Another limitation of this work is that the database was recorded from only one subject. We are collecting similar data from other subjects to validate and expand these results. As suggested by [27], head motion is speaker-dependent. By studying and learning interpersonal differences, head motion sequences can be used to provide a desired personality to the animation [29].

One interesting extension of this work is to include key frame controls to specify gestures such as head "nod" and "tilt" (similar to [16]). Therefore, the designer can add believable head motion in response to facial conversation signals, as proposed by Cassell *et al.* [10].

Another interesting question is how to include other facial gestures such as eyebrow and lip motion in the animation. As mentioned before, gestures and speech are related in a non-trivial manner. Furthermore, different gestures are also related with each other, since most of the time the same set of muscles jointly trigger them. Our research efforts are focused on modeling these relationships to generate facial animations that are perceived to be more natural and engaging.

Acknowledgment

This research was supported in part by funds from the NSF (through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army and a MURI award from ONR. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] I. Albrecht and J. Haber H.P. Seidel. Automatic generation of non-verbal facial expressions from speech. In *Computer Graphics International (CGI 2002)*, pages 283–293, Bradford, United Kingdom, July 2002.

[2] K.S. Arun, T.S. Huang, and S.D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9(5):698–700, September 1987.

[3] P. Boersma and D. Weeninck. Praat, a system for doing phonetics by computer. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, 1996. http://www.praat.org.

[4] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1999)*, pages 21–28, New York, NY, USA, 1999.

[5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH 1997)*, pages 353–360, Los Angeles, CA,USA, August 1997.

[6] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086, March 2007.

[7] C. Busso, Z. Deng, U. Neumann, and S.S. Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, July 2005.

[8] C. Busso and S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech and Language Processing*, In Press, March 2007.

[9] C. Busso and S.S. Narayanan. Interplay between linguistic and affective goals in facial expression during emotional utterances. In *7th International Seminar on Speech Production (ISSP 2006)*, pages 549–556, Ubatuba-SP, Brazil, December 2006.

[10] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In *Computer Graphics (Proc. of ACM SIGGRAPH'94)*, pages 413–420, Orlando, FL,USA, 1994.

[11] E. Chuang and C. Bregler. Mood swings: expressive speech animation. *ACM Transactions on Graphics*, 24(2):331–347, April 2005.

[12] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In *Magnenat-Thalmann N., Thalmann D. (Editors), Models and Techniques in Computer Animation, Springer Verlag*, pages 139–156, Tokyo, Japan, 1993.

[13] R. Cowie and R.R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, April 2003.

[14] D. DeCarlo, C. Revilla, M. Stone, and J.J. Venditti. Making discourse visible: coding and animating conversational facial displays. In *Computer Animation (CA 2002)*, pages 11–16, Geneva, Switzerland, June 2002.

[15] Z. Deng, M. Bulut, U. Neumann, and S. Narayanan. Automatic dynamic expression synthesis for speech animation. In *IEEE 17th International Conference on Computer Animation and Social Agents (CASA 2004)*, pages 267–274, Geneva, Switzerland, July 2004.

[16] Z. Deng, C. Busso, S. Narayanan, and U. Neumann. Audio-based head motion synthesis for avatar-based telepresence systems. In *ACM SIGMM 2004 Workshop on Effective Telepresence (ETP 2004)*, pages 24–30, New York, NY, 2004. ACM Press.

[17] Z. Deng, J.P. Lewis, and U. Neumann. Automated eye motion using texture synthesis. *IEEE Computer Graphics and Applications*, 25(2):24–30, March/April 2005.

[18] Z. Deng, J.P. Lewis, and U. Neumann. Synthesizing speech animation by learning compact speech co-articulation models. In *Computer Graphics International (CGI 2005)*, pages 19–25, Stony Brook, NY, USA, June 2005.

[19] Z. Deng, U. Neumann, J.P. Lewis, T.Y. Kim, M. Bulut, and S. Narayanan. Expressive facial animation synthesis by learning speech co-articultion and expression spaces. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 12(6):1523–1534, November/December 2006.

[20] D. Eberly. *3D Game Engine Design: A Practical Approach to Real-Time Computer Graphics*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2000.

[21] P. Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, April 1993.

[22] P. Ekman and E.L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, New York, NY, USA, 1997.

[23] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. Visual prosody: Facial movements accompanying speech. In *Proc. of IEEE International Conference on Automatic Faces and Gesture Recognition*, pages 396–401, Washington, D.C., USA, May 2002.

[24] B. Granström and D. House. Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46(3-4):473–484, July 2005.

[25] J. Gratch and S. Marsella. Lessons from emotion psychology for the design of lifelike characters. *Applied Artificial Intelligence*, 19(3-4):215–233, March-April 2005.

[26] D. Heylen. Challenges ahead head movements and other social acts in conversation. In *Artificial Intelligence and Simulation of Behaviour (AISB 2005), Social Presence Cues for Virtual Humanoids Symposium*, page 8, Hertfordshire, United Kingdom, April 2005.

[27] H. Hill and A. Johnston. Categorizing sex and identity from the biological motion of faces. *Current Biology*, 11(11):880–885, June 2001.

[28] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936.

[29] L.N. Jefferies, J.T. Enns, S. DiPaola, and A. Arya. Facial actions as visual cues for personality. *Computer Animation and Virtual Worlds*, 17(3-4):371–382, July 2006.

[30] K. Kakihara, S. Nakamura, and K. Shikano. Speech-to-face movement synthesis based on HMMS. In *IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pages 427–430, New York, NY, USA, April 2000.

[31] S. Kettebekov, M. Yeasin, and R. Sharma. Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE Transactions on Multimedia*, 7(2):234–242, April 2005.

[32] T. Kuratate, K. G. Munhall, P. E. Rubin, E. Vatikiotis-Bateson, and H. Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *Sixth European Conference on Speech Communication and Technology, Eurospeech 1999*, pages 1279–1282, Budapest, Hungary, September 1999.

[33] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan. An articulatory study of emotional speech production. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 497–500, Lisbon, Portugal, September 2005.

[34] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, Jan 1980.

[35] Maya software, Alias Systems division of Silicon Graphics Limited. http://www.alias.com, 2005.

[36] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137, February 2004.

[37] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January 1996.

[38] R. W. Picard. Affective computing. Technical Report 321, MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, November 1995.

[39] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[40] M.E. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, and A.M. Tekalp. Combined gesture-speech analysis and speech driven gesture synthesis. In *IEEE International Conference on Multimedia and Expo (ICME 2006)*, pages 893–896, Toronto, ON, Canada, July 2006.

[41] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, April 2003.

[42] K. Shoemake. Animating rotation with quaternion curves. *Computer Graphics (Proceedings of SIGGRAPH85)*, 19(3):245–254, July 1985.

[43] K. Smid, I.S. Pandzic, and V. Radman. Autonomous speaker agent. In *IEEE 17th International Conference on Computer Animation and Social Agents (CASA 2004)*, pages 259–266, Geneva, Switzerland, July 2004.

[44] E. Vatikiotis-Bateson, K.G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia. Characterizing audiovisual information during speech. In *Fourth International Conference on Spoken Language Processing (ICSLP 96)*, volume 3, pages 1485–1488, Philadelphia, PA, USA, October 1996.

[45] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data*, pages 265–268, Kloster Seeon, Bavaria, Germany, May 2000.

[46] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998.

[47] S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan. An acoustic study of emotions expressed in speech. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, 2004.

[48] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England, December 2006.