# Natural Head Motion Synthesis Driven by Acoustic Prosodic Features

Carlos Busso
Dept. of EE
busso@usc.edu

Zhigang Deng
Dept. of CS
zdeng@usc.edu

Ulrich Neumann
Dept. of CS
uneumann@usc.edu

Shrikanth Narayanan
Dept. of EE
shri@sipi.usc.edu [*]

Integrated Media Systems Center
Viterbi School of Engineering
University of Southern California, Los Angeles
http://sail.usc.edu

## Abstract

Natural head motion is important to realistic facial animation and engaging human-computer interactions. In this paper, we present a novel data-driven approach to synthesize appropriate head motion by sampling from trained *Hidden Markov Models* (HMMs). First, while an actress recited a corpus specifically designed to elicit various emotions, her 3D head motion was captured and further processed to construct a head motion database that included synchronized speech information. Then, an HMM for each discrete head motion representation (derived directly from data using vector quantization) was created by using acoustic prosodic features derived from speech. Finally, first order Markov models and interpolation techniques were used to smooth the synthesized sequence. Our comparison experiments and novel synthesis results show that synthesized head motions follow the temporal dynamic behavior of real human subjects.

**Keywords:** Head motion synthesis, prosody, HMM, facial animation, data-driven, spherical cubic interpolation

## 1 Introduction

The development of new human-computer interfaces and exciting applications such as video games and animated feature films has motivated the computer graphics community to generate realistic avatars with the ability to replicate and mirror natural human behavior. Since the use of large motion capture datasets is expensive, and can be only applied to delicately planned scenarios, new automatic systems need to be used to generate natural human facial animation. One useful and practical approach is to synthesize animated human faces driven by speech.

The straightforward use of speech in facial animation is in lip motion synthesis, in which the acoustic phonemes are used to generate visual visemes that match the spoken sentences. Examples of these approaches include [1, 2, 3, 4, 5, 6, 7]. Also, speech has been used to drive human facial expression, under the assumption that the articulation of the mouth and jaw modify facial muscles, producing different faces poses. Examples of these approaches are [8, 9]. Surprisingly, few efforts have focused on natural generation of rigid head motion, which is an important ingredient for realistic facial animations. In fact, Munhall et al. [10] reported that head motion is important for auditory speech perception, which suggests that appropriate head motion can significantly enhance human-computer interfaces.

[*] This is a preprint of an article accepted for publication in Computer Animation and Virtual Worlds Copyright ©(2005) (copyright owner as specified in the journal)

Although human head motion is associated with many factors, such as speaker style, idiosyncrasies and affective states, linguistic aspects of speech play a crucial role. Kuratate et al. [9] presented preliminary results about the relationship between head motions and acoustic prosodic features. They concluded based on the strong correlation ($r$=0.8) that these two are somehow correlated, but perhaps under independent control. This suggests that the tone and the intonation of the speech provide important cues about head motion and vice versa [10]. Notice that, here, it is more important *how the speech is uttered* rather than just *what is said*. Therefore, prosodic features (e.g. pitch and energy) are more suitable than vocal tract-based features (e.g. LPC and MFCC). The work of [11] even reports that about 80% of the variance observed in the pitch can be determined from head motion.

In this paper, an innovative technique is presented to generate natural head motion directly from acoustic prosodic features. First, vector quantization is used to produce a discrete representation of head poses. Then, a *Hidden Markov Model* (HMM) is trained for each cluster, which models the temporal relation between the prosodic features and the head motion sequence. Given that the mapping is not one to one, the observation probability density is modeled with a mixture of Gaussians. The smoothness constraint is imposed by defining a bi-gram model (first order Markov model) on head poses learned from the database. Then, given new speech material, the HMM, working as a sequence generator, produces the most likely head motion sequences. Finally, a smoothing operation based on spherical cubic interpolation is applied to generate the final head motion sequences.

## 2 Related Work

Researchers have presented various techniques to model head motion. Pelachaud et al. [12] generated head motions from labeled text by predefined rules, based on Facial Action Coding System (FACS) representations [13]. Cassell et al. [14] automatically generated appropriate non-verbal gestures, including head motion, for conversational agents, but their focus was only the “nod” head motion. Graf et al. [15] estimated the conditional probability distribution of major head movements (e.g. nod) given the occurrences of pitch accents, based on their collected head motion data. Costa et al. [8] used *Gaussian Mixture Model* (GMM) to model the connections between audio features and visual prosody. The connection between eyebrow movements and audio features was specifically studied in their work. Chuang and Bregler [16] presented a data-driven approach to synthesize novel head motion corresponding to input speech. They first acquired a head motion database indexed by pitch values, then a new head motion sequence was synthesized by choosing and combining best-matched recorded head motion segments in the constructed database. Deng et al. [17] presented a new audio-driven head motion synthesis technique that synthesized appropriate head motion with keyframing control. After a audio-head motion database was constructed, given novel speech input and user controls (e.g. specified key head poses), a guided dynamic programming technique was used to generate an optimal head motion sequence that maximally satisfies both speech and key frames specifications.

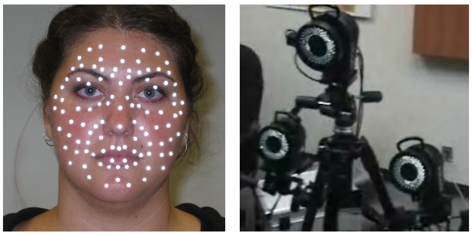

Figure 1: Audio Visual Database

In this paper, we propose to use HMMs to capture the close temporal relation between head motions and acoustic prosodic features. Also, we propose an innovative two-step smoothing technique based on bi-gram models, learned from data, and spherical cubic interpolation.

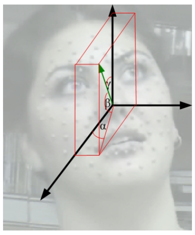

Figure 2: Head poses using Euler angles

# 3 Data Capture and Processing

## *3.1 Database*

The audiovisual database used in this work was recorded from an actress, with 102 markers on her face (left of Fig. 1). She was asked to read a custom, phoneme-balanced corpus four times, expressing different emotions (happiness, sadness, anger and neutral state). A VICON motion capture system with three cameras (right of Fig. 1) was used to capture her facial expressions and head motions. The sampling frequency was set to 120Hz. The recording was made in a quiet room using a close talking SHURE microphone at the sampling rate of 48 kHz. The markers' motions and the aligned audio were captured by the system simultaneously. In total, 633 sentences were used in this work. Note that the actress did not receive any instruction about how to move her head.

After the motion data were captured, all the markers were translated to make a nose marker at the local coordinate center of each frame. A neutral pose was chosen as a reference frame, which was used to create a 102 × 3 matrix, $y$. For each frame, a matrix $x_i$ was created, using the same marker order as the reference. After that, the *Singular Value Decomposition*(SVD), $UDV^T$, of matrix $y^T x_i$ was calculated. Finally, the product of $VU^T$ gave the rotation matrix, $R$, which defines the three Euler angles of the head motion of this frame [18] (Fig. 2).

$$y^T x_i = UDV^T \tag{1}$$

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

To extract the prosodic features, the acoustic signals were processed by the *Entropic Signal Processing System* (ESPS), which computes the pitch (F0) and the RMS energy of the audio. The window was set to 25-ms with an overlap of 8.3-ms. Notice that the pitch takes values only in voiced region of the speech. Therefore, to avoid zeros in unvoiced regions, a cubic spline interpolation was applied in those regions. Finally, the first and second derivatives of the pitch and the energy were added to incorporate their temporal dynamics.

## *3.2 Canonical Correlation analysis*

To validate the close relation between head motion and acoustic prosodic features, as suggested in [9], *Canonical Correlation Analysis* (CCA) [19] was applied to our audiovisual database. CCA provides a scale-invariant optimum linear framework to measure the correlation between two streams of data with different dimensions. The basic idea is to project both feature vectors into a common dimensional space, in which Pearson's correlation can be computed.

Using pitch, energy and their first and second derivatives (6D feature vector), and the angles that define the head motions (3D feature vector), the average correlation computed from the audiovisual database is $r$=0.7. This result indicates that useful and meaningful information can be extracted from the prosodic features of speech to synthesize the head motion.

# 4 Modeling Head Motion

In this paper, we use HMMs, because they provide a suitable and natural framework to model the temporal relation between acoustic prosodic features and head motions. HMMs are used to generate the most likely head motion sequences based on the given observation (prosodic features). The HTK toolkit is used to build the HMMs [20].

The output sequences of the HMMs cannot be continuous, so a discrete representation of head motion is needed. For this purpose, the Linde-Buzo-Gray vector Quantization (LBG-VQ) algorithm [21] is used to define $K$ discrete head poses, $V_i$. The 3D-space defined by the Euler angles is split into $K$ Voronoi regions (Fig. 3). For each region, the mean vector $U_i$ and the covariance matrices $\Sigma_i$ are estimated. The pairs $(U, \Sigma)$ define the finite and discrete set of code vectors

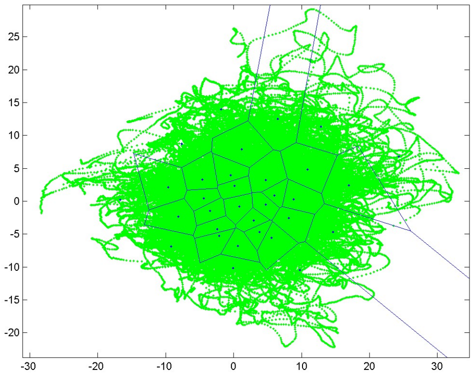

Figure 3: 2D projection of Voronoi regions using 32-size vector quantization

called codebook. In the quantization step, the continuous Euler angles of each frame are approximated with the closest code vector in the codebook.

For each of the clusters, $V_i$, an HMM model will be created. Consequently, the size of the codebook will determine the number of HMM models.

### 4.1 *Learning Natural Head Motion*

The posterior probability of being in cluster $V_i$, given the observation $O$, is modeled according to Bayes rule as

$$P(V_i/O) = c \cdot P(O/V_i) \cdot P(V_i) \tag{3}$$

where $c$ ia a normalization constant. The likelihood distribution, $P(O/V_i)$, is modeled as a Markov process, which is a finite state machine that changes state at every time unit according to the transition probabilities. A first order Markov model is assumed, in which the probabilistic description includes only the current and previous state. The probability density function of the observation is modeled by a mixture of $M$ Gaussian densities which handle, up to some extent, the many-to-many mapping between head motion and prosodic features. Standard algorithms (Forward-backward, Baum-Welch re-estimation) are used to train the parameters of the HMMs, using the training data [22, 20]. Notice that the segmentation of the speech according to the head poses clusters is known. Therefore, the HMMs were initialized with this known alignment (force alignment was not needed).

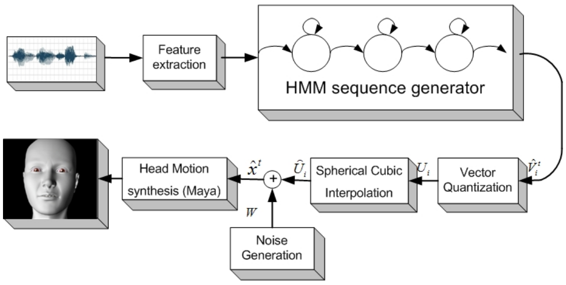

Figure 4: Head motion synthesis framework

The prior distribution, $P(V_i)$, is used to impose a first smoothing constraint to avoid sudden changes in the synthesized head motion sequence. In this paper, $P(V_i)$ is built using bi-gram models, which are learned from data (similar to standard bi-gram language models used to model word sequence probabilities [20, 23]). The bi-gram model is also a first order state machine, in which each state models the probability of observing a given output sequence (in this case, a specific head pose cluster, $V_i$). The transition probabilities are computed using the frequency of their occurrences. In our training database, the inter-cluster transitions are counted and stored, and the statistic learned is used to reward transitions according to their appearances. Therefore, in the decoding step, this prior distribution will penalize transitions that did not appear in the training data.

### 4.2 *Synthesis of Head Motion*

Figure 4 describes the procedure to synthesize head motion. For each testing sample, the acoustic prosodic features are extracted and used as input to the HMMs. The model will generate the most likely sequence, $\widehat{V} = (\widehat{V}_i^t, \widehat{V}_j^{t+1} \ldots)$, where $\widehat{V}_i^t$ is defined by $(U_i, \Sigma_i)$. The mean vector $U_i$ will be used to synthesize the head motion sequences.

The transitions between clusters will introduce breaks in the synthesized signal, even if their cluster means are close (see Fig. 5). Therefore, a second smoothing step needs to be implemented, to guarantee continuity of the synthesized head pose sequences. A simple solution is to interpolate each Euler angle separately. However, it has been shown that this technique is not optimal, because it introduces jerky movements and other undesired effects such as *Gimbal lock* [24]. As suggested by Shoemake, a better approach is to interpolate in the quaternion unit sphere [24]. The basic idea is to transform the Euler angles into quaternions, which are an alternative rotation matrix representation, and then interpolate the frames in this space.

In this paper, we used spherical cubic interpolation [25], squad, which is based on spherical linear interpolation, slerp. For two quaternions $q_1$ and $q_2$, the slerp function is defined as:

$$\mathrm{slerp}(q_1, q_2, \mu) = \frac{\sin(1-\mu)\theta}{\sin\theta} q_1 + \frac{\sin\mu\theta}{\sin\theta} q_1 \quad (4)$$

where $cos\theta = q_1 \cdot q_2$ and $\mu$ are variables that range from 0 to 1 and determine the frame position of the interpolated quaternion. Given four quaternions, the squad function is defined as:

$$\begin{aligned} &\mathrm{squad}(q_1, q_2, q_3, q_4, \mu) = \qquad (5) \\ &\quad \mathrm{slerp}(\mathrm{slerp}(q_1, q_4, \mu), \mathrm{slerp}(q_2, q_3, \mu), 2\mu(1-\mu)) \end{aligned}$$

After the Euler angles are transformed into quaternions, key-points are selected by downsampling the quaternions at a rate of 6 frames per second (this value was empirically chosen). Then, spherical cubic interpolation is used in those key-points by using the squad function. After interpolation, the frame rate of the quaternions is 120 frames per second, as the original data. The last step in this smoothing technique is to transform the interpolated quaternions into Euler angles. Figure 5 shows the interporlation result for one of the sentences. The resulting vectors are denoted $\widehat{U}_i$. The readers are referred to [25] for further details about spherical cubic interpolation.

Finally, the synthesized head pose, $\hat{x}^t$, at time $t$ will be estimated as:

$$\hat{x}^t = (\alpha, \beta, \gamma)^T = \widehat{U}_i + W_i \quad (6)$$

where $W_i$ is a zero-mean uniformly distributed random white noise. Notice that $\hat{x}^t$ is a blurred version of $V_i$'s mean.

If the size of the codebook is large enough, the quantization error will be insignificant. However, the number of HMMs needed will increase and the discrimination between classes will decrease. Also, more data will be needed to train the models. Therefore, there is a tradeoff between the quantization error and the inter-cluster discrimination.

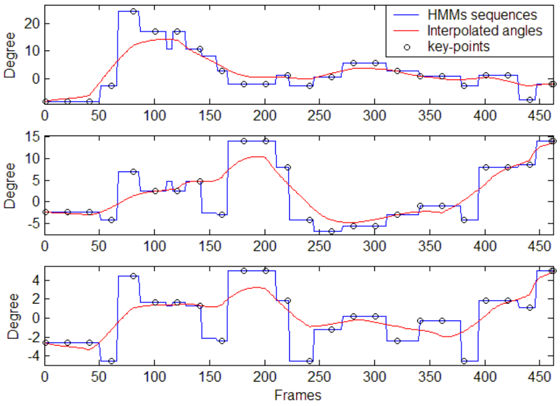

Figure 5: Spherical cubic interpolation

### 4.3 *From Euler Angles to Talking Avatars*

A blend shape face model composed of 46 blend shapes is used in this work (eye ball is controlled separately). To create a realistic avatar, lip and eye motion techniques are also included. For novel text/audio input, the speech animation approach presented in [6, 26] was used to generate synchronized visual speech. Then, eye motion is automatically synthesized by a texture-synthesis based approach [27]. Hence, appropriate blend shape weights are calculated for each frame. After that, the same audio is used to generate corresponding natural head motion using the proposed approach. The generated head motion Euler angles, $\hat{x}^t$, are directly applied to the angle control parameters of the face model. This animation procedure is applied to each frame. Besides the animation, the face modeling and rendering are also done in Maya.

## 5 Results and Discussion

### 5.1 *HMM configuration*

The topology of the HMM is defined by the number and the interconnection of the states. The most popular configurations 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 transitions between all the states are allowed. The LR topology is simple and needs less data to train its parameters. The EG topology is less restricted, so it can learn a larger set of state transitions from the data. In this particular problem, it is not clear which topology gives better description of the head motion dynamics. Therefore, eight HMM configurations, described in table 1, with different topologies, number of models, $K$, number of states, $S$, and number of mixtures, $M$, were trained. Notice that the size of the database is not big enough to train more complex HMMs with more states, mixtures or models than those described in table 1.

As mentioned before, the pitch, the energy and their first and second derivatives are used as acoustic prosodic features to train each of the proposed HMM configuration. Eighty percent of the database was used for training and twenty percent for testing.

### 5.2 *Objetives evaluation*

To evaluate the performance of our approach, the prosodic features from the test data were used to generate head motion sequences, as described in previous section. For those samples, the Euclidian distance, $d_{euc}$, between the Euler angles of the original frames and the Euler angles of the synthesized data, $\hat{x}^t$, was calculated. The average and the standard deviation of all the frames of the testing data, is shown in table 1 ($\overline{D}$). Notice that the synthesized head motions are directly compared with the original data (not its quantized version), so the quantization error is included in the values of the table. As mentioned in the introduction, head motion does not depend only on prosodic features, so this level of mismatch is expected.

Table 1 also shows the canonical correlation analysis between the synthesized and original data. As can be observed the correlation was around $r$=0.85 for all the topologies. This result strongly suggests that the synthesized data follow the behavior of the real data, which validates our approach.

As can be seen from table 1, the performance of the different HMM topologies are similar. The Left-to-Right HMM, with a 16-size codebook, 3 states and 2 mixtures achieves the best result. However, if the database were big enough, an ergodic topology with more states and mixture could perhaps give better results. The next experiments were implemented using this topology (K=16 S=3 M=2 LR).

| ***HMM config.*** | $\overline{D}$ | | CCA | |
|---|---|---|---|---|
| | Mean | Std | Mean | Std |
| K=16 S=5 M=2 LR | 10.2 | 3.4 | 0.88 | 0.11 |
| K=16 S=5 M=4 LR | 9.3 | 3.4 | 0.87 | 0.11 |
| K=16 S=3 M=2 LR | 9.1 | 3.6 | 0.87 | 0.12 |
| K=16 S=3 M=2 EG | 9.1 | 3.4 | 0.87 | 0.10 |
| K=16 S=3 M=4 EG | 9.5 | 3.4 | 0.83 | 0.12 |
| K=32 S=5 M=1 LR | 12.8 | 4.0 | 0.83 | 0.14 |
| K=32 S=3 M=2 LR | 10.7 | 3.3 | 0.86 | 0.12 |
| K=32 S=3 M=1 EG | 10.4 | 3.1 | 0.86 | 0.11 |

Table 1: Results for different configurations

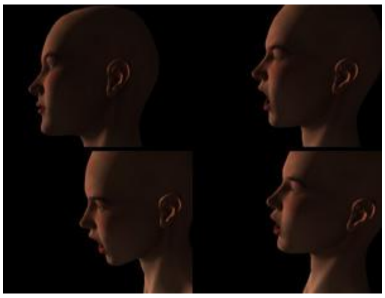

Figure 6: Synthesized data, side view

### 5.3 *Head motion animation results*

We also recorded sentences, not included in the corpus mentioned before, to synthesize novel head motion using our approach. For each recorded audio, the procedure described in this paper was applied. Figures 6 and 7 show four frames of the synthesized data. For animation results, please refer to the accompanying video.

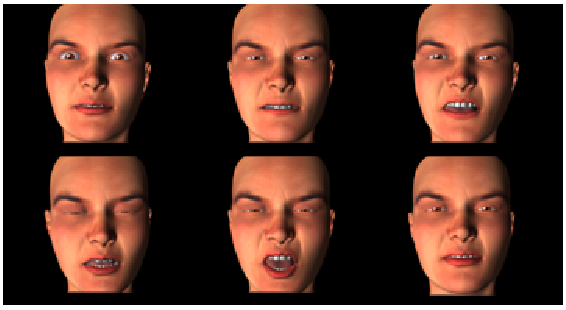

Figure 7: Synthesized data, front view

# 6 Conclusions

This paper presents a novel approach to synthesize natural human head motions driven by speech prosody. HMMs are used to capture the temporal relation between the acoustic prosodic features and head motions. The use of bi-gram models in the sequence generation step guarantees smooth transitions from the discrete representations of head movement configurations. Furthermore, spherical cubic interpolation is used to avoid breaks in the synthesized signal.

The results show that the synthesized sequences follow the temporal dynamic behavior of real data. This proves that the HMMs are able to capture the close relation between speech and head motion. The results also show that the smoothing techniques used in this work can produce continuous head motion sequences, even when only a 16 word sized codebook is used to represent head motion poses.

In this paper we show that natural head motion animation can be synthesized by using just speech. In future work, we will add more components to our system. For example, if the emotion of the subject is known, as is usually the case in most of the applications, suitable models that capture the emotional head motion pattern can be used, instead of a general model.

## Acknowledgments

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 under Contract EEC-9529152, a CAREER award (to S. Narayana), and the Department of the Army under Contract DAAD 19-99-D-0046. Special Thanks go to J.P. Lewis and Murtaza Bulut for helping data capture, Hiroki Itokazu, Bret St. Clair, and Pamela Fox for face model preparation, and anonymous reviewers for useful comments and suggestions.

## References

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

[2] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. *Proceedings of ACM SIGGRAPH'97,* pages 353–360, 1997.

[3] M. Brand. Voice puppetry. *Proceedings of ACM SIGGRAPH'99,* pages 21–28, 1999.

[4] S. Kshirsagar and N. M. Thalmann. Visyllable based speech animation. *Computer Graphics Forum (Proc. of Eurographics'03),* 22(3), 2003.

[5] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transaction on Graphics(Proceedings of ACM SIGGRAPH'02),* pages 388–398, 2002.

[6] 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.

[7] W. Liu, B. Yin, and X. Jia. Audio to visual signal mappings with hmm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04),* pages 885–8, Quebec, Canada, May 2004.

[8] M. Costa, T. Chen, and F. Lavagetto. Visual prosody analysis for realistic motion synthesis of 3d head models. In *International Conference On Augmented, Virtual Environments and Three Dimensional Imaging,ICAV3D,* Ornos, Mykonos, Greece, May-June 2001.

[9] T. Kuratate, K. G. Munhall, P. E. Rubin, E. V. 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.

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

[11] H. Yehia, T. Kuratate, and E. V. Bateson. Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data*, pages 265–268, 2000.

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

[13] P. Ekman and W. V. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Printice-Hall, 1975.

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

[15] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang. Visual prosody: Facial movements accompanying speech. In *Proc. of IEEE Int'l Conf. on Automatic Faces and Gesture Recognition*, 2002.

[16] E. Chuang and C. Bregler. Head emotion. *Stanford University Computer Science Technical Report, CS-TR-2003-02*, 2003.

[17] 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.

[18] M. B. Stegmann and D. D. Gomez. A brief introduction to statistical shape analysis, mar 2002.

[19] C. Dehon, P. Filzmoser, and C. Croux. Robust methods for canonical correlation analysis. In *Data Analysis, Classification, and Related Methods*, pages 321–326, Springer-Verlag, Berlin, 2000.

[20] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England., 2002.

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

[22] 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.

[23] X. Huang, A. Acero, and H-W. Hon. *Spoken Language Processing: A guide to theory, algorithm and system development*. Printice-Hall, 2001.

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

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

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

[27] 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.