Natural Head Motion Synthesis Driven by Acoustic Prosodic Features

Carlos Busso, Zhigang Deng, Ulrich Neumann, Shrikanth Narayanan

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

Oct 18th, 2005

Overview

- Motivation
- Data Capture and Processing
- Modeling Head Motion
- Results and Discussion
- Conclusion

Motivation

- Engaging human-computer interfaces and application such as animated features films have motivated realistic avatars
- A useful and practical approach is avatars driven by speech
- Straightforward use of speech: lip motion (vocal tract features) [Liu, 2004] [Ezzat, 2002]
- Head motion and prosodic features are closely related [Kuratate, 1999]
  - Correlation between head motion and prosodic features .83
  - Motion of the head is integrated with the system that generate speech, but under independent control

Motivation

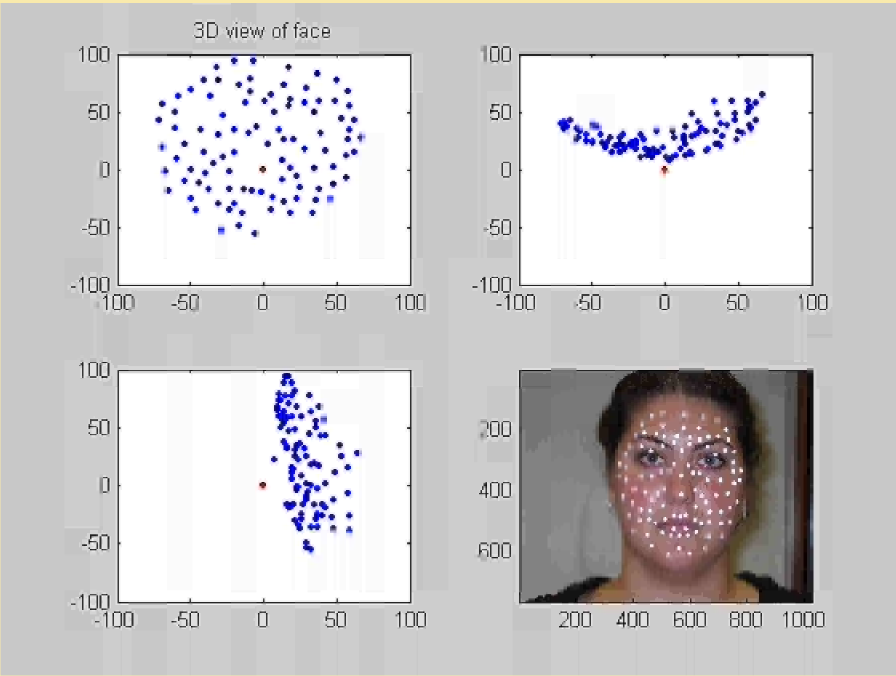
- Further evidence
  - Head motion is important for auditory speech perception [Munhall,2002]
  - 80% of the variance of the pitch can be determined from head motion [Yehia, 2000]

- Proposed framework
  - *Hidden Markov Models* are trained capture the temporal relation between the prosodic features and the head motion sequence
  - Vector quantization is used to produce a discrete representation of head poses
  - Two-step smoothing technique based on first order Markov model and spherical cubic interpolation

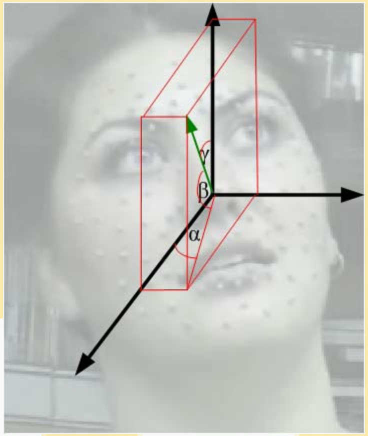
- Previous Work
  - Rule-based systems: [Pelachaud, 1994]
  - Gaussian Mixtures Model [Costa, 2001]
  - Specific head motion (e.g. ‘nod’) [Cassell, 1994] [Graf, 2002]
  - Example-based system [Deng, 2004], [Chuang, 2004]

Data Capture and Processing

- Database
  - An actress read 633 utterances expressing different emotions (angry, happy, sad and neutral)
  - Video:
    - Sample rate: 120 fps
    - VICON capture system
    - Head Motion features (α,β,γ) extracted with SVD [Stegmann, 2002]
  - Audio:
    - Sample rate: 48 KHz
    - Window: 25 ms
    - Overlap: 8.3 ms
    - Pitch and RMS energy extracted using ESPS



Data Capture and Processing



- Features
  - Head Pose: 3 angles (α,β,γ) (3D features vector)
  - Audio: Pitch, RMS energy and their first and second derivative (6D feature vector)
- Canonical Correlation Analysis
  - Scale-invariant optimum linear framework to measure the correlation between two streams of data with different dimensions [Dehon, 2000]
  - The average correlation computed from the audiovisual database (Head poses vs. prosodic feature) is r=0.7
  - Useful and meaningful information can be extracted from the prosodic features to synthesize the head motion

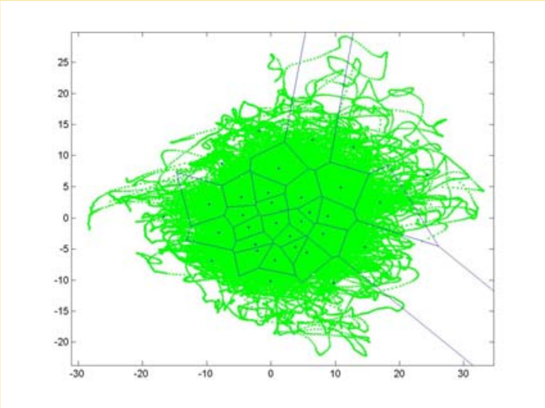
Modeling Head Motion

- Head motion are modeled with HMMs
  - HMMs provide a suitable and natural framework to model the temporal relation between acoustic prosodic features and head motions
  - HMMs will be used as sequence generator (head motion sequence)

- Discrete head pose representation
  - The 3D head motion data is quantized using K-dimensional vector quantization

$$HeadPose = (\alpha, \beta, \gamma) \approx V_i \qquad i \in \{1..K\}$$

  - Each cluster is characterized by its mean, U_i, and covariance, Σ_i



Modeling Head Motion

- Learning Natural Head motion
  - $P(V_i | O) = c \cdot P(O | V_i) P(V_i)$
  - The observation, *O*, are the acoustic prosodic features
  - One HMM will be trained for each head pose cluster, V_i

Likelihood distribution $P(O | V_i)$

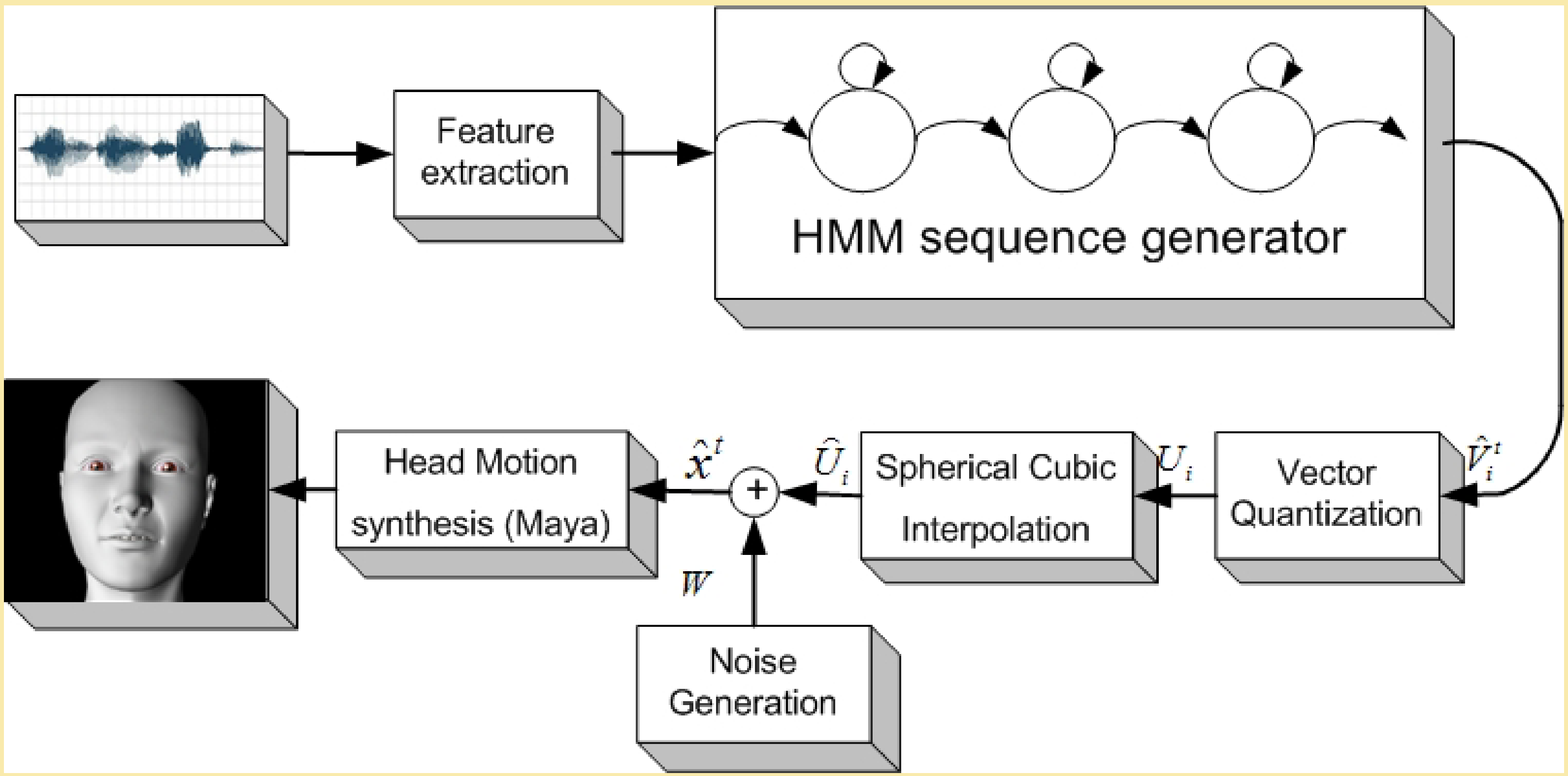
- It is modeled as a Markov process
- A mixture of *M* Gaussian densities is used to model the *pdf* of the observations
- Standard algorithm are used to train the parameters (Forward-backward, Baum-Welch re-estimation)

Prior distribution $P(V_i)$

- It is built as bi-gram models learned from the data (1st smoothing step)
- Transitions between clusters that do not appear in the training data are penalized
- This smoothing constraint is imposed in the decoding step

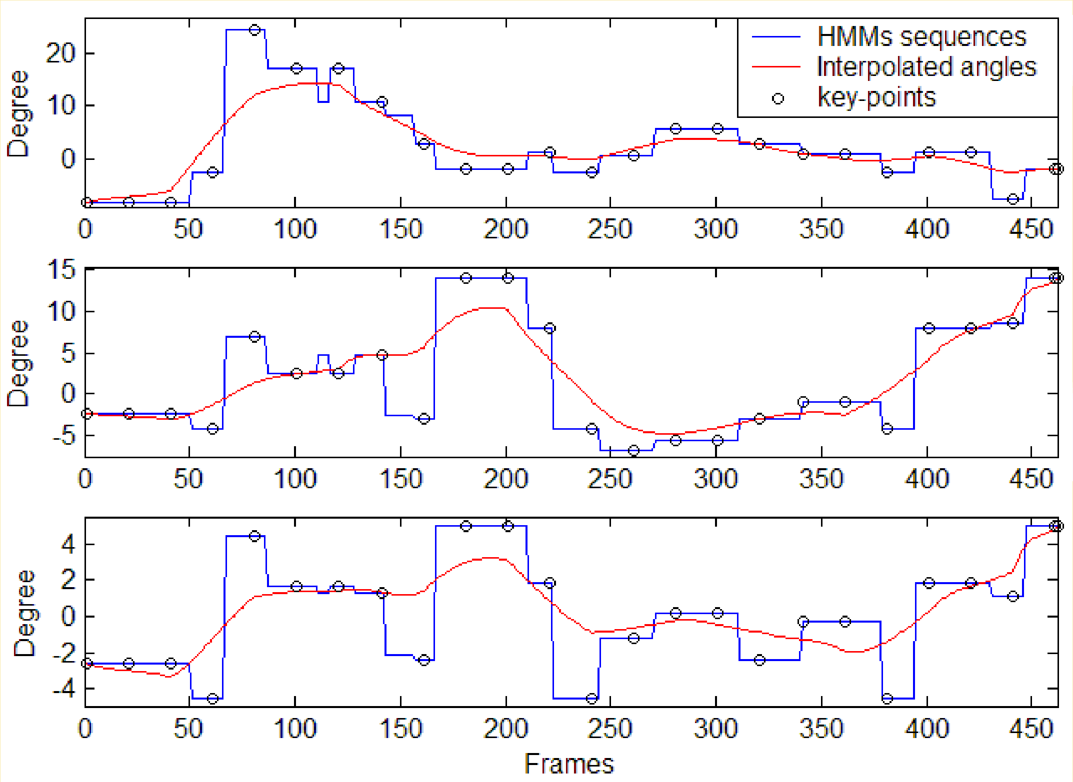
Modeling Head Motion

- Synthesis of head motion
  - For a novel sentence, the HMMs generate the most likely head motion sequence
  - Interpolation is used to smooth the cluster transition region (2nd smoothing step)



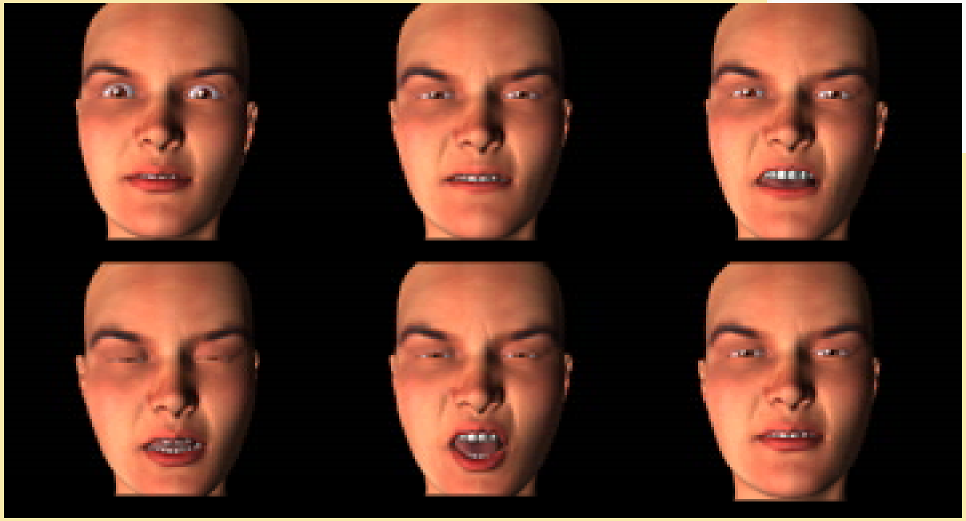
Modeling Head Motion

- Spherical Cubic Interpolation
  - 2nd smoothing constraint
  - Remove the breaks in the cluster transition of the generated sequences
  - The interpolation take place in the quaternion unit sphere [Shoemake, 1985]



Modeling Head Motion

- From Euler Angles to Talking Avatars
  - Avatar is synthesized using Maya
  - A model with 46 blend shapes is used
  - Lip and eye motions are also included [Deng, 2004][Deng, 2005] [Deng_2, 2005]
  - The Euler angles are directed applied to the control parameters of the face model



Results and Discussion

- HMM configuration
  - Eight HMM configurations were used
    - *K*, number of cluster (number of models)
    - S, number of states
    - M, number of mixtures
    - LR, Left-to-Right topology
    - EG, Ergodic topology
  - Eighty percent of the database is used for training and twenty percent for testing
- Objective evaluation
  - Euclidean distance and Canonical Correlation Analysis between the real head motion sequence and the synthesized data

Results and Discussion

D, Euclidean Distance
CCA, Canonical correlation analysis
K, number of cluster (number of models)
S, number of states
M, number of mixtures
LR, Left-to-Right topology
EG, Ergodic topology

- Objective evaluation (cont.)

HMM config.	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

- Synthesized data follow the temporal pattern of real head motion (r=0.85)
- There is a expected mismatch between the real and synthesized data
  - Head motion depend also on other factors (speaker style, idiosyncrasies, emotions)

Results and Discussion

- Head motion animation results
  - Sequence 1: Speech from same subject of training data
  - Sequence 2: Speech from another subject

Conclusion

- General observation
  - Speech prosody provides enough information to synthesize realistic avatars
  - The synthesized sequences follow the temporal dynamic behavior of real data
  - The HMMs are able to capture the close relation between speech and head motion
  - 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.
- Future work
  - Use HMMs for each emotion instead of global models
  - Include eyebrows, which also have strong correlation with prosodic features
  - Use a different discrete representation of head poses

Spherical Cubic Interpolation

- Interpolation procedure
  - Euler angles are transform to quaternion
  - Key-points are selected by down-sampling the quaternion sequence
  - Spherical cubic interpolation (squad) is used to interpolate those key-points
  - The interpolated results are transformed to Euler angles

$$squad(q_1, q_2, q_3, q_4, u) = slerp(slerp(q_1, q_4, u), slerp(q_2, q_3, u), 2u(1-u))$$

$$slerp(q_1, q_2, u) = \frac{\sin(1-u)\theta}{\sin\theta} q_1 + \frac{\sin u\theta}{\sin\theta} q_2$$

- Motivation for spherical cubit interpolation
  - Interpolation in Euler space introduce jerky movement
  - Introduce undesired effects such as Gimbal lock