DATA-DRIVEN SYNTHESIS OF EXPRESSIVE VISUAL SPEECH USING AN MPEG-4 TALKING HEAD

Jonas Beskow & Mikael Nordenberg, 2006

presented by Andrew Sabatino
ajs2136@columbia.edu
October 26, 2011
EVALUATION OF THE EXPRESSIVITY OF A SWEDISH TALKING HEAD IN THE CONTEXT OF HUMAN-MACHINE INTERACTION

Jonas Beskow & Loredana Cerrato, 2005
Synthetic Expressive Visual Speech

- Applications in web services, automated tutors, virtual avatars in computer games, animation
- Emotions modeled independently using principal components and coarticulated alongside synthetic speech
Develop emotionally indexed corpus of expressions

- 75 short sentences
- Each rendered with happiness, sadness, surprise, disgust, fear, anger, and neutral
- 525 total utterances
- 29 IR-sensitive markers attached to speaker’s face
  - 4 reference markers on ears and forehead
  - Setup conforms to MPEG-4 feature point (FP) configuration
  - Sub-millimeter accuracy
  - 60 frames/second
- Sync signal fed into digital video and audio track
Marker placement on the face of a Swedish, amateur actor
Talking Head

- MPEG-4 facial animation standard
- Textured, 3D, male face with 15K polygons
- Face controlled by facial animation parameters (FAPs), corresponding to a number of feature points on the face.
- 68 FAPs are specified in MPEG-4, but 38 relevant FAPs were used in this work
  - Relevancy was empirically determined
- FAPs expressed in normalized FAPUs (Facial Animation Parameter Units) which take the distances between facial landmarks to normalize across specific facial models
Control of the Head

- Face is controlled by deformations applied to the FAPs.
- Features like head and eye rotations use rotational deformation.
- Other deformations are centered on one vertex:
  - Surface distance to surrounding, outer vertices is calculated.
  - Weighting function:
    - \( w = \text{vertex weight}, \ d = \text{edge distance}, \ r = \text{FAP-specific influence radius} \)
    - \( w = e^{-\frac{d^2}{r^2}} \)
Snapshots taken at 0.1 s from “I will buy…” (in Swedish)

1. Happy  
2. Angry  
3. Surprised  
4. Sad
Cohen-Massaro model of coarticulation

- Each phonetic segment has a target vector of articulatory parameters
- These are smoothed over time using exponential growth and decay of parameters with slopes that can be adjusted for each parameter
- As such, parameters can be said to have a trajectory with different coefficients for growth and decay
- This study minimizes error between predicted and measured trajectories of individual features
  - Prior work had set the constants empirically
\[ z(t) = \frac{\sum_{i=1}^{N} T_i D_i(t)}{\sum_{i=1}^{N} D_i(t)} \]

**Figure 2** Cohen-Massaro model of a three-segment utterance

\[ D_i(t) = \begin{cases} 
A_i e^{-\theta_i (\tau_i - t) c_i} & t < T_i \\
A_i e^{-\phi_i (\tau_i - t) c_i} & t \geq T_i 
\end{cases} \]

Ti – target value

Di(t) – dominance function
Principal component analysis (PCA) performed to reduce co-dependency between adjacent points on face and therefore neighboring FAPs. Top 10 principal components explain 99% of FAP variation in acted data stream. 5 separate models trained: happy, sad, angry, surprised, neutral using Gauss-Newton minimization function \texttt{fminunc} in Matlab iteratively until the error stopped decreasing. Of about 70 sentences for each emotion, 10 used for testing and rest used for training. Control models created and later integrated with articulation algorithms to produce novel Swedish speech.
Tests

- Models used to synthesize animations in sync with synthesized versions of text sentences.
- For 10 test subjects with 4 categories: happy, angry, sad, or neutral:
  - 73% happy accurately recognized
  - 60% angry accurately recognized
  - 40% sad accurately recognized
- Viable eyebrow motions arose from training, were especially effective for anger.