Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head

Jonas Beskow & Mikael Nordenberg

KTH Speech, Music and Hearing
SE-10044 Stockholm, Sweden
beskow@speech.kth.se

Abstract

This paper describes initial experiments with synthesis of visual speech articulation for different emotions, using a newly developed MPEG-4 compatible talking head. The basic problem with combining speech and emotion in a talking head is to handle the interaction between emotional expression and articulation in the orofacial region. Rather than trying to model speech and emotion as two separate properties, the strategy taken here is to incorporate emotional expression in the articulation from the beginning. We use a data-driven approach, training the system to recreate the expressive articulation produced by an actor while portraying different emotions. Each emotion is modelled separately using principal component analysis and a parametric coarticulation model. The results so far are encouraging but more work is needed to improve naturalness and accuracy of the synthesized speech.

1. Introduction

In recent years, there has been an increased interest for animated characters in a diverse array of applications: web services, automated tutors for e-learning, avatars in virtual environments, and computer games. Further, the concept of embodied conversational agents (ECAs) -- animated agents that are able to interact with a user in a natural way using speech, gesture and facial expression -- holds the potential of a new level of naturalness in human-computer interaction, where the machine is able to convey and interpret verbal as well as non-verbal communicative acts, ultimately leading to more robust, efficient and intuitive interaction.

Audio-visual speech synthesis, i.e. production of synthetic speech with properly synchronised movement of the visible articulators, is an important property of such agents that not only improves realism, but also adds to the intelligibility of the speech output [1]. Previous work on visual speech synthesis has typically been aimed at modelling neutral pronunciation. However, as the agents and the systems they embody become more advanced, the need for affective and expressive speech arises. This presents a new challenge in acoustic as well as in visual speech synthesis. Several studies have shown how articulation is affected by expressiveness in speech, in other words, articulatory parameters behave differently under the influence of different emotions ([2], [3]). This interdependency between emotional expression and articulation has made it difficult to combine simultaneous speech and emotional expression in synthetic talking heads.

This paper describes recent experiments with synthesis of expressive emotional speech articulation in a virtual talking head. Rather than trying to model speech and emotion as two separate properties, the strategy has been to incorporate emotional expression in the articulation from the beginning. We have used a data-driven approach, training the system to recreate the expressive articulation produced by an actor while portraying different emotions.

2. Synthesis of Visible Speech

Visual speech synthesis can be accomplished either through manipulation of video images ([4], [5]) or based on two- or three dimensional models of the human face and/or speech organs that are under control of a set of deformation parameters, as described by for example [6], [7], [8] and [9].

To render visual speech movements, we start from a time-aligned transcription of the speech to be synthesized. The time-aligned transcription can be obtained from a text-to-speech system, if we are synchronising with synthetic speech, or it can be produced by a phoneme recognizer (as in the Synface system [10]) or a phonetic aligner [11].

Next, we need an articulatory control model, i.e. an algorithm that can convert a time-aligned phonetic transcription of the utterance into control parameter trajectories to drive the articulation of the talking head model. In order to produce convincing and smooth articulation, the articulatory control model will have to model coarticulation, which refers to the way in which the realisation of a phonetic segment is influenced by neighbouring segments. Different strategies have been proposed to deal with coarticulatory effects in visual speech synthesis, based on rules, speech production theories or machine learning algorithms.

In a recent study [12] several different articulatory control models were implemented and their performance was evaluated. The models were data-driven, and trained on a set of sentences spoken with neutral articulation and recorded using 3D motion capture. Two of the models were based on theoretical models of co-articulation, and two were based on artificial neural networks. Each of the models was automatically trained by adjustment of free parameters in order to minimize the error between prediction and the measured trajectories. Evaluation was done by comparing the ability of the different models to accurately predict measured trajectories, as well as through an audiovisual intelligibility study. In this study it was concluded that while there were no significant differences between the different data-driven models in the intelligibility study, the models differed slightly in how well they were able to predict the measured trajectories. The model that produced the lowest overall error was the Cohen-Massaro coarticulation model [7], making it a natural choice also for the present task of modelling expressive articulation.

Thus, in the present study, with the goal of performing the same kind of training, but with training data representing emotional articulation, The Cohen-Massaro model was chosen.

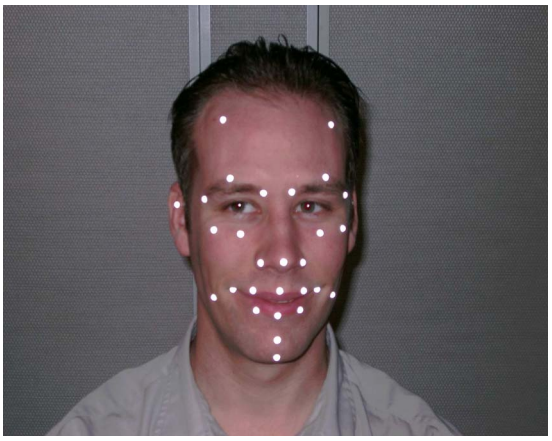

Figure 1: Marker placements on the speakers face for the recording of the 75 sentences used to train the models for expressive articulation.

3. Data Collection

We have used an opto-electronic motion tracking system: Qualysis MacReflex – to collect a multimodal corpus of acted emotional speech. The Qualisys system allows capturing the dynamics of emotional facial expressions, by registering the 3D coordinates of a number of reflective markers, with sub-millimetre accuracy, at a rate of 60 frames/second. For this study, our speaker, a male native Swedish amateur actor, was instructed to produce 75 short sentences with the six emotions happiness, sadness, surprise, disgust, fear and anger, plus neutral, yielding 7 x 75 recorded utterances.

A total of 29 IR-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers (on the ears and on the forehead). The marker setup (as shown in figure 1) largely corresponds to MPEG-4 feature point (FP) configuration. Audio data was recorded on DAT-tape, and video was recorded using a mini-DV digital video camera. A synchronisation signal from the Qualisys system was fed into one audio channel of the DAT and DV to facilitate post-synchronisation of the data streams.

4. MPEG-4 Face Animation

Our talking head is based on the MPEG-4 Facial Animation standard [13]. It is a textured 3D-model of a male face comprised of approximately 15000 polygons. The mesh has been parameterised to allow for realistic deformation, using a framework based around a combination of professional 3D-modelling tools and in-house custom algorithms, and a flexible animation engine.

The MPEG-4 standard allows the face to be controlled directly by number of parameters (FAPs, facial animation parameters). The FAPs specify the movements of a number of feature points in the face. The full set consists of 68 FAPs, but in this study a sub-set of 38 FAPs were used since many were not considered relevant[1]. FAPs are expressed in normalized units called FAPUs (Facial Animation Parameter Units) which are defined by distances between facial landmarks. This causes FAPs to be independent of the specific face model, which is an important property of the MPEG-4 specification, making it is possible to drive a face model from points measured on a face that differs in geometry with respect to the model.

4.1. FAP calculation

In order to utilise the 3D point registrations described in the previous section, the data first had to be converted into MPEG-4 FAPs. First the marker movements were decomposed into global and local movement. This was done by expressing all points in a local coordinate system that is fixed to the head, which is consistent with MPEG-4 definition of axes: x-axis pointing left, y-axis pointing upward and z-axis pointing straight forward, in the direction of the nose. This coordinate system was defined using the reference markers on the ears and upper forehead. Global head rotation angles were also calculated, but not utilised in the present experiment.

Given that marker placement in the recording session corresponded to MPEG-4 feature points, calculation of the FAP values is achieved using linear relations and normalisation factors. If marker coordinates are arranged in a matrix X where each row represents a time frame and each column a coordinate for one of the markers, then a corresponding FAP matrix F can be calculated as

$$F = F_0 + M \cdot U \cdot (X - X_0) \quad (1)$$

Where X_0 represents the markers in resting position, for which corresponding FAP values F_0 have been manually estimated. M is a (manually constructed) matrix that maps marker coordinates (columns in X) to FAPs (columns in F), and U is a diagonal matrix containing the proper FAPU scaling factors for each FAP.

4.2. Animation

Our face is controlled by the FAPs using a number of deformations. FAPs controlling rotations, like head and eye rotations, use rotational deformations, while the other FAPs use weighted deformations. The weighted deformations employ a technique known as skinning, which means that it applies a deformation matrix to each affected vertex, where each vertex has its own weight value specifying how much it should be influenced by that deformation. The FAPs having weighted deformations are associated to a feature point specifying the centre of the deformations, as described in the MPEG-4 specification. Applying weights to the areas around the centre feature points is a three-step process. For each feature point we first calculate an approximation of the surface distance between each vertex and the feature point. Then we map this distance to a weight value using the following weighting function:

$$w = e^{\frac{-d^2}{r^2}} \quad (2)$$

where w is the vertex weight, d is the edge distance and r is a FAP specific influence radius.

A prerequisite for extracting the individual FAP values directly from the tracked points of the recordings, as described in the previous section, is that deformations from different

[1] In particular, the special "high-level" FAPs termed *viseme* and *expression*, intended to directly define visual speech targets and prototypical emotions respectively, were not used, since they allow for neither co-articulation nor expressive articulation. FAPs dealing with eye gaze direction, nose, ears and global head rotation were also omitted.

Figure 2: Snapshots taken every 0,1 s from the animation of the the fragment "*jag ska köpa…*" ("*I will buy…*"), synthesized with four different expressive speech models: happy (top row), angry (2nd row), surprised (3rd row) and sad (bottom row).

FAPs are mutually independent. The last step is therefore to adjust the weight maps to make deformations from different FAPs independent of each other.

5. Modelling and Synthesis of Expressive Articulation

The recorded corpus with expressive speech was used to train articulatory control models for five of the recorded emotions: happy, angry, surprised, sad and neutral. These control models can later be used to synthesize articulatory movements for novel (arbitrary) Swedish speech, thereby modelling expression and articulation in an integrated fashion.

5.1. Articulatory control model

We have adopted the coarticulation model by Cohen & Massaro (1993). In the Cohen-Massaro model, each phonetic segment is assigned a target vector of articulatory parameters. The target values are then blended over time using a set of overlapping temporal dominance functions. The dominance functions take the shape of a pair of negative exponential functions, one rising and one falling. The height of the peak, the rate with which the dominance rises and falls, as well as the shape of the slope (exponent) are free parameters that can be adjusted independently for each phoneme and articulatory control parameter.

The trajectory of a parameter $z(t)$ can be calculated as

$$z(t) = \frac{\sum_{i=1}^{N} T_i D_i(t)}{\sum_{i=1}^{N} D_i(t)} \qquad (3)$$

where N is the number of segments in the utterance, T_i denotes the target value for segment i and $D_i(t)$ denotes the dominance function for segment i, which given by the equation

$$D_i(t) = \begin{cases} \mathrm{A}_i \cdot e^{-\theta_i(\tau_i - t)^{c_i}} & t < \mathrm{T}_i \\ \mathrm{A}_i \cdot e^{-\varphi_i(t - \tau_i)^{c_i}} & t \geq \mathrm{T}_i \end{cases} \qquad (4)$$

where τ_i is the centre time of the segment, α_i is a scaling factor used to control the degree of dominance for each segment, θ_i and φ_i are coefficients for forward- and backward coarticulation respectively.

In it's original application by Cohen and Massaro, the model's free parameters were empirically determined through hand-tuning and repeated comparisons between synthesis and video recordings of a human speaker.

In this work, we apply a data-driven training procedure based on minimisation of the error between predicted and measured trajectories.

5.2. Principal Components

In the Cohen-Massaro algorithm, each articulatory parameter is modelled independently, thus it is advantageous if the modelled parameters can be considered reasonably independent. However, there is a considerable co-dependency between adjacent points in the face, and thus between MPEG-4 FAPs. One way to reduce co-dependency while at the same time decreasing the number of parameters is to perform a principal component analysis (PCA). For each of the emotions in the corpus, a separate PCA was performed. The top 10 principal components were able to explain 99% of the variation in the original FAP data streams.

5.3. Training

The data was phonetically labelled based on the audio files using an automatic forced alignment procedure [11], and the phonemes were mapped to 25 viseme categories.

Each of the top 10 PC's were modelled individually using the Cohen-Massaro model of coarticulation. Five separate models were trained, one for each of the emotions happy, sad, angry, surprised and neutral). Of the approximately 70 sentences available for each emotion (a few were discarded due to measurement problems), ten were set aside for testing, and the rest were used for training. The training was carried out using the Matlab function `fminunc`, which implements the Gauss-Newton minimisation algorithm, with the error function defined as the summed squared distance between the measured and the predicted tracks. The training was terminated when the error over the test set stopped decreasing. To increase speed of convergence and robustness of the training, gradient information was also taken into account. See [12] for a detailed description of the training procedure including calculation of the gradients.

The resulting models can be used to generate PC trajectories, which are converted back to FAP trajectories by multiplication with the matrix of PCA basis vectors for the specific emotion. Thus it is possible to create expressive visual speech synthesis with the desired emotion to accompany arbitrary acoustic speech, either natural (phonetically aligned) or synthetic.

6. Evaluation

The resulting models were used to synthesize animations of a number of the sentences available in the test set, in order to facilitate comparisons between directly recorded material and the synthesised variants. The subjective impression was that the most convincing emotions were also the most extreme ones: happy and angry, while sad and surprised were more subtle but also more difficult to distinguish. Figure 2 shows snapshots from resulting animations of the same utterance for the four emotions happy, angry, surprised and sad.

To quantify the perceived emotion, a diagnostic perceptual experiment was conducted. This experiment aimed at identification of emotion from stimuli where different synthesized visual emotions were combined with neutral acoustic speech. The task was to classify the perceived expression into one of four categories: happy, angry, sad or neutral. The average recognition rate (for 10 subjects) was 73% for happy, 60% for angry and 40% for sad, which is well above chance level. The full experiment, which was set in the context of a language training application, is reported in [14].

7. Conclusions

The work presented in this paper should be considered as a first attempt at synthesising expressive visual speech. While still not perfect, it seems clear from both the informal and the perceptual evaluation that the proposed technique is quite effective. An advantage with the approach is that not only articulation is modelled, but the movements of the whole face is included - eyebrows, cheeks etc. An unexpected side effect of this was that plausible eyebrow movements clearly emerged from the training, especially for the angry model, even though the only input to the control model was phonemes. At the same time there are certain problems with the current models, extreme movements especially for angry and happy, sometimes result in unrealistic facial configurations. Our next steps in this area will be to include contextual information in the training data, and investigate alternative control models, followed by a more thorough perceptual evaluation.

8. Acknowledgements

This work was carried out in the framework of the EU-project PF-Star, funded by the European Commission, proposal number: IST2001 37599. Thanks also Bertil Lyberg for providing access to the Qualisys measurement facility.

9. References

[1] Siciliano C, Williams G, Beskow J & Faulkner A. "Evaluation of a Multilingual Synthetic Talking Face as a communication Aid for the Hearing Impaired", in *Proceedings of ICPhS 2003*, Barcelona, Spain, Aug. 2003 pp. 131-134.

[2] Nordstrand, M., Svanfeldt G., Granström B., House D. Measurements of articulatory variation in expressive speech for a set of Swedish vowels. *Journal of Speech Communication* 44, 2004, pp.187-196.

[3] Magno Caldognetto E., Cosi P., Cavicchio F., "Modifications of Speech Articulatory Characteristics in the Emotive Speech". *Proc. Affective Dialogue Systems 2004*, pp. 233-239

[4] Bregler, C, Covell, M. and Slaney, M. "Video Rewrite: Driving Visual Speech with Audio." *Proceedings of ACM SIGGRAPH'97*, 1997, pp. 353-360.

[5] Ezzat, T, Geiger, G., Poggio, T. "Trainable Videorealistic Speech Animation". *Proceedings of ACM SIGGRAPH'02, San Antonio, TX*, 2002, pp. 388-398.

[6] Beskow, J. "Animation of Talking Agents" *Proceedings of International Conference on Auditory-Visual Speech Processing*, Rhodos, Greece, 1997, pp. 149-152

[7] Cohen, M. M. and Massaro, D. W. "Modelling Coarticulation in Synthetic Visual Speech" Magnenat-Thalmann N., Thalmann D. (Eds), *Models and Techniques in Computer Animation*, Springer Verlag, Tokyo, 1993, pp. 139-156.

[8] Cosi, P., Fusaro, A., Grigoletto, D., & Tisato, G. "Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes". *ADS 2004*: 101-112

[9] Pelachaud, C. "Visual Text-to-Speech" In Pandzic, I. & Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, John Wiley & Sons, 2002, pp. 125-140.

[10] Beskow J, Karlsson I, Kewley J and Salvi G. "SYNFACE - A Talking Head Telephone for the Hearing-impaired." In K Miesenberger, J Klaus, W Zagler, D Burger eds *Computers helping people with special needs*, 2004, pp. 1178-1186

[11] Sjölander K and Heldner M "Word level precision of the NALIGN automatic segmentation algorithm". *Proc Fonetik 2004* 116-119

[12] Beskow, J. "Trainable Articulatory Control Models for Visual Speech Synthesis", *Journal of Speech Technology* 7(4), 2004, pp. 335-349.

[13] Pandzic, I. S. and Forchheimer, R. *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons, 2002

[14] Beskow, J. & Cerrato, L. "Evaluation of The Expressivity of a Swedish Talking Head In The Context of Human-Machine Interaction", *Proceedings of GSCP '04*, Padova, Italy, 2005.