Expressive Animated Agents for Affective Dialogue Systems

Jonas Beskow, Loredana Cerrato, Björn Granström, David House, Mikael Nordenberg, Magnus Nordstrand and Gunilla Svanfeldt[1]

KTH Speech Music and Hearing, 100 44 Stockholm, Sweden
{beskow, loce, bjorn, david, mikaeln, magnusn, gunillas}@speech.kth.se

Abstract

We present our current state of development regarding animated agents applicable to affective dialogue systems. A new set of tools are under development to support the creation of animated characters compatible with the MPEG-4 facial animation standard. Furthermore, we have collected a multimodal expressive speech database including video, audio and 3D point motion registration. One of the objectives of collecting the database is to examine how emotional expression influences articulatory patterns, to be able to model this in our agents. Analysis of the 3D data shows for example that variation in mouth width due to expression greatly exceeds that due to vowel quality.

Introduction

At KTH we have for a long time been developing animated talking heads and studying their use in various human-machine [1,2] and human-human settings [3]. This paper presents an overview of the current status of a number of activities related to the area of animated agents that are applicable to affective speech-based dialogue systems. The activities include development of new tools and models for expressive facial animation and collection of multimodal corpora of expressive speech. These activities are partly conducted within the EU-IST project PF-STAR, a project aiming at defining technological baselines within several areas related to affective interaction with machines, for example emotional speech synthesis and facial animation.

Face modelling

We are developing a new set of tools and techniques to facilitate the development and animation of face models adhering to the MPEG-4 Facial Animation standard [4]. The standard defines 66 low-level facial animation parameters (FAPs) that describe

[1] Names in alphabetic order

the animation of a face model. Compared to previous facial animation work at KTH [5], that has been based around a compact articulatory motivated parameter scheme, the MPEG-4 FA provides finer detail in the specification of facial expression. Another advantage of the standard is that it opens up the possibility of sharing data and models with other researchers in the area, a fact that is central to the work within the PF-STAR project, where there are three different sites involved in facial animation work, and the MPEG-4 standard has been chosen as the common format for data exchange.

Development of an animated face model is typically a tedious and time-consuming task, often relying on custom made tools for geometry manipulation and parameterisation. The current goal is to develop tools that automate the model creation process, enabling creation of high-quality standards compliant MPEG-4 FA models from arbitrary static 3D-meshes with a minimum of user-intervention. Furthermore, we wish to take advantage of the capabilities available in state-of-the-art 3D-modelling packages when it comes to actual sculpting and texturing of the models. We have chosen to make our tools compatible with such a 3D-modelling package called XSI[2]. The tools consist of three main parts: a plug-in to XSI for selecting landmark points on the face, a MATLAB script to build the animated model, and a custom rendering engine. In order to create a new animated model, the model constructor will first create or import a static facial mesh in XSI, and then use the plug-in to identify a number of landmark points on the mesh, known as MPEG-4 Facial Definition Points or FDPs for short [4]. Once the FDPs have been identified, a heuristically based approach is used to automatically obtain the weights that dictate to what degree each individual vertex in the mesh is influenced by a given facial animation parameter (FAP). The resulting model can then be animated by the rendering engine, which may be embedded into applications such as dialogue systems.

To increase the expressiveness of the model, the rendering engine includes the capability of dynamically render wrinkles in the face based on local estimates of the compression of the skin [6]. To achieve real-time performance, we use hardware-accelerated bump-mapping, leveraging the power of the latest generation of graphics processing units (GPUs).

Data recording and re-synthesis

To gain knowledge about how to drive our agents, in terms of expressive non-verbal and verbal behaviour, we have collected multimodal corpora of emotive speech using an opto-electronic motion tracking system, *MacReflex* from Qualisys[3]. By using reflective markers applied on the speaker's face it is possible to record the 3D position for each marker with sub-millimetre accuracy, every 1/60th second, by using four infrared cameras. 35 markers were used to record lip movements as well as other facial movements such as eyebrows, cheek, chin and eyelids. Five markers attached to

[2] http://www.softimage.com

[3] http://www.qualisys.se

a pair of spectacles were used as a reference to be able to factor out head and body movements. In addition to 3D marker positions, video and audio was recorded.

Two corpora of expressive speech have been collected. *Corpus 1* was a sample recording aimed at evaluating the feasibility of different elicitation techniques such as reading prompts and interactive dialogue. *Corpus 2* consisted of non-sense words and short sentences, providing good phonetic coverage.

Corpus 1 was made up of two sub-corpora, one of prompted speech and one of naturally elicited dialogues. The prompted material consisted of digit sequences and semantic neutral utterances. 15 different expressions were chosen; together with the six universal prototypes for emotions: *anger*, *fear*, *surprise*, *sadness*, *disgust* and *happiness* [7], we also had the subject to act *worried*, *satisfied*, *insecure*, *confident*, *questioning*, *encouraging*, *doubtful*, *confirming* and *neutral*. For the dialogue sub-corpus, an information-seeking scenario was used. This communicative scenario is similar to the one that might arise between a user and an embodied conversational agent in a dialogue system. One of the dialogue participants had the role of "information giver". The domains were movie information (plots, schedules), and direction giving. The focus of the recording was on the "information giver", and only his movements were recorded. Audio recordings included both subjects.

Corpus 2 consisted of VCV & VCCV nonsense words, CVC nonsense words and short sentences. An actor read the words and sentences while acting six different emotional states, a sub-set of the emotions used in corpus 1: *confident, confirming, questioning, insecure, happy, neutral*. A total of 1700 items were recorded.

In order to apply the recorded 3D data to the face models, MPEG-4 FAPs were extracted, by establishing linear relationships between the MPEG-4 FAPs and the displacements of the markers from the neutral state (after compensating global head motion).

One limitation of the optical motion tracking systems when applied to articulatory measurement is that it is not possible to place markers on the inner lip contour, since the markers would obstruct the subject's articulation. Therefore, a pre-processing stage was performed prior to FAP extraction, where points measured on the outer lip contour were used to estimated points on the inner contour, using a linear estimator, as described in [8].

Expressive articulation

Most systems for visual speech synthesis are modelled on non-expressive speech, i.e. the material is read with a neutral voice and facial expression. However, expressiveness might affect articulation and how we produce speech a great deal, and an articulatory parameter might behave differently under the influence of different emotions. This can be deduced from a quick analysis of vowels in our database, of which an example is presented in figure 1. The mean position of the left mouth corner measured in the middle of all the vowels in the material is displayed as a cross, the size of one standard deviation.

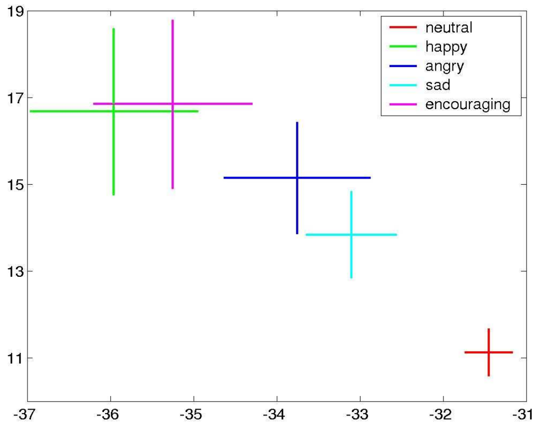

Fig. 1. Mean horizontal and vertical position of the left mouth corner for five acted expressive states: *happy*, *encouraging*, *angry*, *sad* and *neutral* from left to right.

It can be seen that the expressive state in some instances has a stronger influence on the articulation than do the different vowels. It is also interesting to note that the neutral pronunciation displays a pattern different from all the (acted) expressive speech versions, with very little variation between vowels and a presumably small mouth opening. In this study we did not look into the dynamic influence on the segmental articulation in the expressive speech. How much could be described by relatively stable settings and what is best described by expressive gestures is the topic of some of our current research.

References

[1] Gustafson J, Lindberg N & Lundeberg M (1999). The August spoken dialogue system. Proc of Eurospeech 99, 1151-1154.

[2] Gustafson, J. Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000). AdApt–a multimodal conversational dialogue system in an apartment domain. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000)*. Bejing, China, pp. 134-137.

[3] Siciliano C, Williams G, Beskow J and Faulkner A (2003). Evaluation of a Multilingual Synthetic Talking Face as a communication Aid for the Hearing Impaired. Proc 15th ICPhS

[4] Ostermann, J. (2002). Face Animation in MPEG-4. In Pandzic, I. S. and Forchheimer, R. (Eds.) *MPEG-4 Facial Animation - the Standard, Implementation and Applications*, Chichester, England: John Wiley & Sons. pp. 17-56.

[5] Beskow, J. (2003). Talking Heads – Models and Applications for Multimodal Speech Synthesis, Doctoral Dissertation, KTH, Stockholm, Sweden.

[6] Nordenberg, M., 2003. *Modelling and rendering dynamic wrinkles in a virtual face*.TMH/KTH MSc thesis. (http://www.speech.kth.se/qpsr/masterproj/)

[7] Ekman P. An Argument for Basic Emotions. In Basic Emotions. N.L. Stein and K. Oatley (eds), pp 169-200, 1992.

[8] Beskow, J.; Engwall, O.; Granström, B., 2003. Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. of ICPhS 2003*. Barcelona, Spain.