

Reactive Memories: An Interactive Talking-Head

Vincent E. Devin and David C. Hogg
School of Computing
University of Leeds
Leeds LS2 9JT, UK

Abstract

We demonstrate a novel method for producing a synthetic talking head. The method is based on earlier work in which the behaviour of a synthetic individual is generated by reference to a probabilistic model of interactive behaviour within the visual domain - such models are learnt automatically from typical interactions. We extend this work into a combined visual and auditory domain and employ a state-of-the-art facial appearance model. The result is a synthetic talking head that responds appropriately and with correct timing to simple forms of greeting with variations in facial expression and intonation.

keywords : **interactive head, face tracking, speech reconstitution, virtual partner, behaviour modelling.**

1 Introduction

The screen-based 'talking head' is a powerful device for mediating interaction between humans and machines, enabling a form of interaction that mimics direct communication between humans [10, 1, 11]. The experience of realism is further enhanced when the computer is equipped with visual and auditory senses with which to perceive the user [6, 2, 13, 4]. In this symmetric situation, both the human and synthetic head can see and be seen, and can hear and be heard.

Of paramount importance within face to face conversation is of course the content of what is said. However, the sequence and timing of accompanying facial expressions is also important; mistimed or inappropriate expressions may convey unintended meaning and can therefore be disruptive. It is reasonable to suppose the same requirements will apply for human interaction with a synthetic talking head.

An approach that begins to meet these requirements is proposed in [9] and [6]. Their idea is based on the common notion of a state space, in which each vector represents the instantaneous configuration of a participant in an interaction. Such vectors are the endpoint of a perceptual process within the computer, sensing the human party, and the start-point for a graphical process generating the synthetic individual. An interaction can be thought of as a pathway through the joint configuration space corresponding to the human and synthetic party in an interaction. The range of possible interactions is represented as a stochastic process over the joint configuration space, which is learnt through observation of real interactions captured on video. Johnson et al. [8] modeled the profiles of two people

shaking hands. Jebara et al. [6] modeled head and hand gestures. In both cases, the models were used to drive a synthetic individual in response to past joint behaviour.

In the current paper, we construct a simple synthetic talking head by adopting the same approach together with combined modeling of speech and facial expression. The principal objective has been to produce a reactive head in which speech utterances and facial expressions are both appropriate and timely. We handle only a few kinds of simple verbal interactions. Nevertheless, the resulting system is indicative of a new kind of medium that could replace the photograph in an album or the home video with a reactive icon of a familiar person - hence the title of the paper.

2 Representing facial appearance and sound

Configurations of the talking head over a sampling interval (typically around 0.05 second) are represented by the parameters of a facial appearance model, based on that proposed in [3], combined with the principal components of spectral coefficients from the corresponding sound fragment. These choices were determined by the need for an internal representation that was both concise, to facilitate construction of a stochastic process model, and capable of being mapped back on to realistic images and sounds.

2.1 Facial appearance

Facial appearance is represented by the parameters of a combined model of shape and intensity variation for the human face [3]. This model is generated from training data of typical faces, marked-up by hand with spline curves delineating prominent facial structures (Figure 1). A principal component analysis (PCA), applied to the parameters of the facial lines within an aligned frame of reference, recovers an underlying set of axes of shape variation.

Any face shape x can then be approximated with :

$$x = \bar{x} + P_s b_s \quad (1)$$

where \bar{x} is the mean shape, P_s is a matrix with columns the principal orthogonal modes of variation and b_s is a vector of shape parameters. Figure 2 illustrates the facial lines generated from three different sets of values for the parameters of such a shape model.

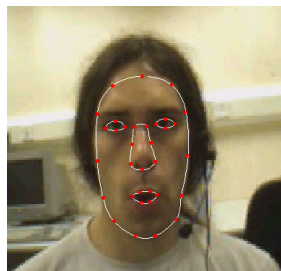


Figure 1: Spline curves delineating prominent structures of the face

A grey-level appearance model is constructed by warping each face from the training set onto the mean shape and applying principal component analysis to the normalised data.

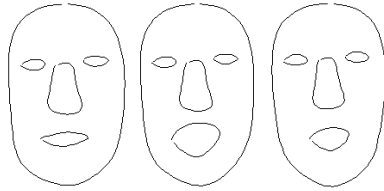


Figure 2: Face shapes generated from three different sets of parameters values

The warping is performed by triangulating between points on each facial line (see figure 3) and applying an affine mapping between corresponding triangles. Any normalized grey-level face g can then be approximated with :

$$g = \bar{g} + P_g b_g \quad (2)$$

with \bar{g} the mean grey-level face, P_g a matrix with columns the principal orthogonal modes of variation and b_g a vector of grey-level parameters.

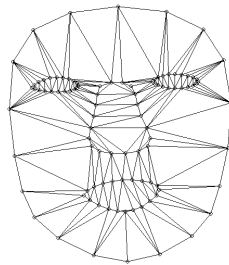


Figure 3: Mean shape triangulated for warping

A face can now be synthesised from shape and grey-level parameters by generating the normalized grey-level image and warping it to the given shape (figure 4)



Figure 4: Faces generated from three sets of values for model parameters

In the experiments reported in this paper, there are 10 parameters in the final facial appearance model to which are added 4 parameters for position, orientation and scaling (giving a total of 14 parameters). A separate model is required for each individual. We do not model variations in identity, although the modeling framework can be extended to do this (see [3]). In general, a specific face is represented by assigning values to the n parameters of the facial appearance model: $\mathbf{f} = \{f_0, f_1, \dots, f_{n-1}\}$

The mapping from a given visual image to the corresponding parameters of the appearance model is performed using the iterative search proposed in [3]. The idea is to adjust model parameters so that the corresponding synthetic face matches the new image as closely as possible. Figure 5 shows a close match between an input image and the synthetic face chosen as best match.

The difference between a new image and one synthesized by the appearance model needs to be minimised. A difference vector δI can be defined:

$$\delta I = I - I_s \quad (3)$$

where I is the vector of grey-level values in the image, and I_s is the vector of synthesized grey-level values for the current parameters. To locate the best match between model and image, the aim is to minimise the magnitude of the difference vector, $\Delta = |\delta I|^2$, by varying the model parameters \mathbf{f} . The relationship between δI and the error in the model parameters is assumed to be linear :

$$\delta \mathbf{f} = A \delta I \quad (4)$$

To find A , multivariate linear regression is performed on a random sample of model displacements $\{\delta \mathbf{f}_i\}$ and the corresponding difference images $\{\delta I_i\}$.



Figure 5: The closest matching synthetic face to a new image

2.2 Sound

The speech waveform is partitioned into frames containing 512 samples. In our experiments, the speech signal is sampled at 11kHz, giving 21.53 frames per second. A spectral analysis is performed on each frame using a fast fourier transform [12].

A principal component analysis on the vectors of spectral components from a training set of utterances allows a concise representation of individual frames. For our experiments, the first 70 principal components were found to give adequate reconstruction of the waveform (figure 6). Thus, each frame (512 waveform samples, lasting 46ms) is represented by 70 parameters. In general, a specific frame of sound is represented by assigning values to the m parameters of the sound model: $\mathbf{s} = \{s_0, s_1, \dots, s_{m-1}\}$

The sound model is able to reproduce with fidelity all the vocabulary occurring in the utterances it has been trained with. Any utterance not occurring in the training set would be expected to give a poor reconstitution, like that shown on the bottom-left of figure 6. In figure 7, the residual error for different utterances with the use of different numbers of

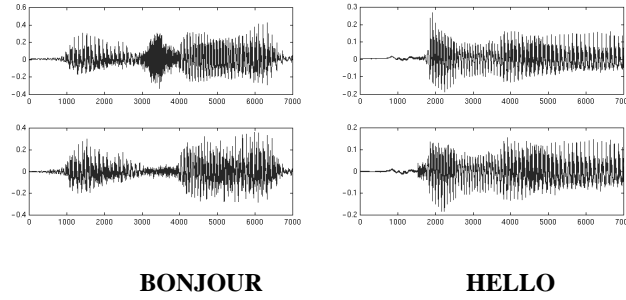


Figure 6: On the left the sound 'bonjour' (top) with reconstitution (bottom). This sound is not in the training set. On the right the sound 'Hello' (top) with reconstitution (bottom)

principal components can be observed. Note that the french 'bonjour' is relatively well matched globally (i.e. a comparable residual) even though it was not part of the training set.

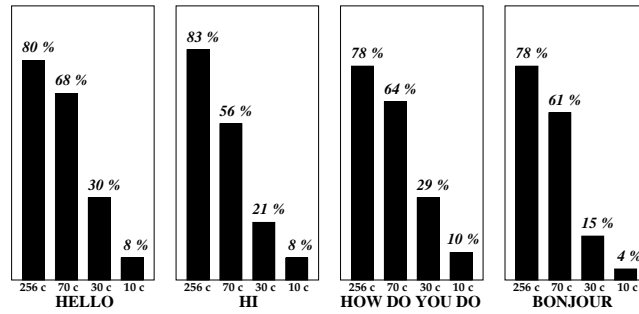


Figure 7: Quality of reconstitution of different utterances with the use of different numbers of components. Only 'Bonjour' did not occur in the training set.

3 Representing interactions

A separate face model and sound model are built for each speaker. In our experiments, we use 15 training sequences of the same pair of individuals, recorded at 15 frames per second for the video and 11kHz for the sound. The video is resampled to 21.53fps to match the rate at which sound frames occur. Facial appearance is encoded in 14 parameters and individual sound frames are encoded in 70 parameters.

An interaction is represented by the joint behaviour of the two individuals involved. At any given instant, the joint configuration is described by a *combined face/speech vector* \mathbf{C}_t (see figure 8). The temporal evolution of an interaction is represented by an ordered set of *state vectors* $\{\mathbf{F}_0, \mathbf{F}_1, \dots, \mathbf{F}_k\}$, consisting of the combined vector \mathbf{C}_t and its scaled first derivative $\lambda \dot{\mathbf{C}}_t$, approximated in the training data by the difference in combined vector between successive frames :

$$\mathbf{F}_t = (\mathbf{C}_t, \lambda \dot{\mathbf{C}}_t) \tag{5}$$

The resultant vector derived from the training data has 168 components.

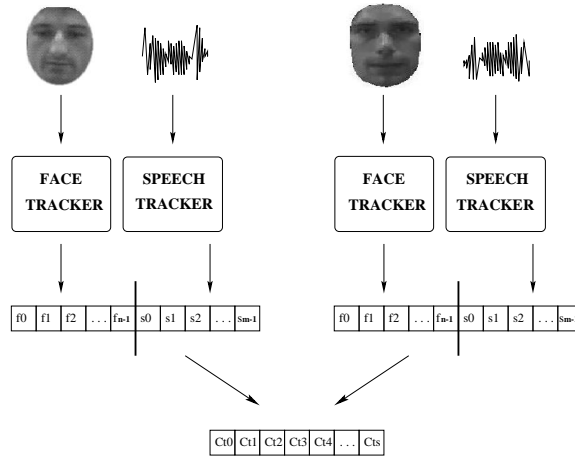


Figure 8: Joint vector of the combined face/speech vectors for the two talking-heads

In [9], a probability density function over the space of possible interactions is learnt from training data of typical interactions - in their case, handshakes viewed in profile. The method used for representing the density function was used earlier in [7] for the characterization of general behaviours which last an indefinite time and involve a high-dimensional state space.

The method is in two parts. First the state vectors derived from the training data are approximated by a set of state prototypes placed by vector quantisation.

The second part builds on the first to provide a vector representation for extended behaviours. The way in which this is achieved may be visualised as follows. Each state prototype has an associated activation level that is initialised to zero for the encoding of a single behaviour. The behaviour is traced from beginning to end, and at each time step the activation of each state prototype either decays by a fixed proportion or takes on a value that is a linearly decreasing function of the prototype's distance from the current state, whichever gives the largest value. The pattern of activation levels at each time-step provides an encoding of the behaviour up until that point on the trajectory and this is recorded as a vector - referred to as a behaviour vector to distinguish from the state vectors.

The same procedure is repeated for all behaviours in the training set, giving a large collection of behaviour vectors encoding these trajectories and all partial trajectories implicit in their generation. This approach to encoding the evolution of a system is equivalent to the so-called leaky neural network model. Finally, the distribution of behaviour vectors extracted from the training data are themselves encoded by a set of prototype behaviours, again derived by vector quantisation.

The final behaviour prototypes provide a compressed model for the range of behaviours observed in the training data. This model can be adapted to serve as a piecewise uniform probability density function in which each prototype is replaced by a uniform region with magnitude proportional to the local density of prototypes, which is in turn proportional to the observed density of training behaviours (for details see [7]).

We adopt the same method for encoding interactions. In our experiments, a set of 500 168-dimensional state prototypes were obtained from vector quantisation in the first part

of the method, and a set of 500 500-dimensional behaviour prototypes were obtained in the second part.

4 Generating a response

Unfortunately, the behaviour model is not easily generative in the sense that it might be used to produce a sample interactive behaviour. Although the information required is contained in the representation, it is not easily extracted. To rectify this problem, a Markov chain is superimposed on top of the behaviour prototypes, with transitions defining the ways in which behaviours in the neighbourhood of prototypes may evolve between time-steps. The probability of a transition is estimated from the proportion of such transitions observed in the training set.

Although this extension to the behaviour model satisfies the Markov property that the probability of moving to the next prototype is dependent only on the current prototype, because the current prototype encodes the history within itself, there remains a longer duration temporal dependence. This is not a Markov chain over state prototypes.

In tracking the user during an interaction, we must deal with uncertainty in the elements of that part of the state vector acquired from pre-processing each incoming frame. A Bayesian framework is adopted in which the posterior density for the hypothesised state \mathbf{F}_t at each time-step is estimated recursively from a prior density for the state and the likelihood function given the current observation \mathbf{S}_t^H :

$$P(\mathbf{F}_t | \mathbf{S}_t^H, \dots, \mathbf{S}_0^H) \propto P(\mathbf{S}_t^H | \mathbf{F}_t) P(\mathbf{F}_t | \mathbf{S}_{t-1}^H, \dots, \mathbf{S}_0^H) \quad (6)$$

where $P(\mathbf{F}_t | \mathbf{S}_t^H, \dots, \mathbf{S}_0^H)$ is the conditional distribution of state given an observation history, $P(\mathbf{S}_t^H | \mathbf{F}_t)$ measures the *likelihood* of a state \mathbf{F}_t , giving rise to observation \mathbf{S}_t^H , and $P(\mathbf{F}_t | \mathbf{S}_{t-1}^H, \dots, \mathbf{S}_0^H)$ is the *prior* distribution representing predictions from the *posterior* distribution $P(\mathbf{F}_{t-1} | \mathbf{S}_{t-1}^H, \dots, \mathbf{S}_0^H)$, from the previous time step.

A Gaussian likelihood function is used, based on the hypothesis' error $E(\mathbf{F}_t, \mathbf{S}_t^H)$:

$$P(\mathbf{S}_t^H | \mathbf{F}_t) = \exp\left(-\frac{E(\mathbf{F}_t, \mathbf{S}_t^H)^2}{2\sigma^2}\right). \quad (7)$$

We implement the CONDENSATION tracking algorithm of Isard and Blake [5], in which the posterior density is represented by a set of sample hypotheses. In our experiments, a total of 100 sample hypotheses were found to be adequate.

In general, the maximum of the posterior density provides a plausible final hypothesis of the current state of an interaction. Unfortunately, it is hard to estimate this from the representation of the density. Instead the state hypothesis that maximises the likelihood function is selected, and the corresponding synthetic facial appearance and sound fragment generated. This turns out to satisfy our purposes in practice.

5 Behaviour filter

The Markov chain describe all the behaviour seen in the training sequences. With a large and varied training data set all the basics *action/reaction* should be modelled. Any se-

quence of a *correct* behaviour should be able to go through the Markov Chain using the maximum likelihood keeping the error $E(\mathbf{F}_t, \mathbf{S}_t^H)$ under a fixed threshold.

On the figure 9 we can observe the error variation when an *incorrect behaviour* is performed

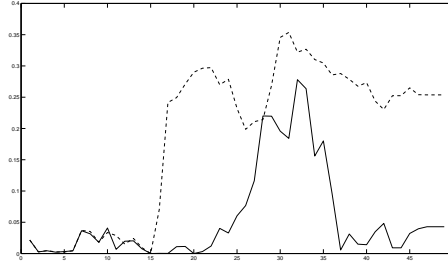


Figure 9: Unusual behaviour in dotted line, usual one in plain line. The vertical axis is the $E(\mathbf{F}_t, \mathbf{S}_t^H)$ value and on the horizontal one the sequence of images

This *behaviour threshold* (BT) can be used as a filter for a correct behaviour. Any incorrect behaviour would be detected when the error $E(\mathbf{F}_t, \mathbf{S}_t^H)$ goes over the threshold.

In a situation when an answer is expected, like a smile or a simple chiming¹ face, the system could take over if nothing happens. The algorithm to go through the Markov Chain is :

1. Select the initial hypothesis \mathcal{H}_0 from the set X_0 of all potential initial hypotheses such that the error $E(\mathbf{F}_0, \mathbf{S}_0^H)$ is minimised.
2. if $E(\mathbf{F}_t, \mathbf{S}_t^H) < BT$ then produce the observed talking response S_t^R else produce the virtual human's response S_t^V from \mathcal{H}_t .
3. Select the future hypothesis \mathcal{H}_{t+1} from the set X_{t+1} of all potential future hypotheses such that the error $E(\mathbf{F}_{t+1}, \mathbf{S}_{t+2}^H)$ is minimised. The potential hypotheses \mathcal{X}_{t+1} are extrapolations at time $t + 1$ from \mathcal{H}_t .
4. Repeat steps 2-3 until the end state is reached.

6 Results

We show results from a set of experiments in which the system is trained with simple interactions involving the greetings 'Hello', 'Hi' and 'How do you do?', with associated responses and facial expressions on both sides (The utterances used are shown in table 1).

Training data is acquired from a pair of cameras and microphone headsets attached directly to workstations at which the two participants in an interaction are seated. Figure 10 shows a single frame from each video stream and the complete speech waveforms for a single interaction from the training set.

At present, the system is too slow (about 2fps) to allow real-time responses to the user. For this reason, optimal responses are generated off-line to a pre-recorded greeting. This

¹Chiming : from the use of 'to chime in'. If someone **chimes in**, they say something just after someone else has spoken, usually to agree with them or to support their argument [Collins Cobuild dct]

Learning sequences			
	Question	Answer	Number
1	“hello ?!”	“Hello !”	5
2	“Hi ?!”	“Hi there !”	4
3	“how do you do ?”	“Fine !”	5
4	-	-	1

Table 1: Learning sequences : different interactions with different intonations

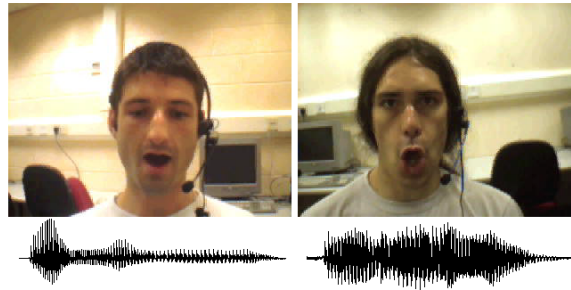


Figure 10: Illustrating a typical interaction from the training set

opens the natural feedback loop in an interaction between two individuals but nevertheless demonstrates correct timing and appropriateness of responses. We do not expect the real-time system (nearing completion) to perform in a qualitatively different way.

The waveform and a single video frame from a pre-recorded greeting is shown on the left in figure 11. The response of the synthetic head is shown alongside this on the right.

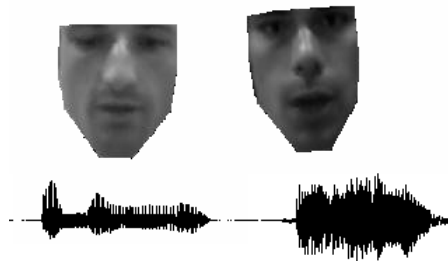


Figure 11: A single speaker saying "hello" is tracked (and the virtual partner answering "hello" is displayed)

7 Conclusion

The results obtained are preliminary and for a non-real-time system. However, the overall framework linking speech with video within a reactive system has been demonstrated. We will shortly have completed a real-time version and this will facilitate more rapid experimentation and closure of the feedback loop between the user and the synthetic head.

The approach presented is not intended to deal with anything but simple forms of interaction of the kind shown. Broader use of English would introduce a new dimension of complexity that is beyond its scope. Nevertheless, we believe the work has taken a step in the direction of the living pictures of dead persons in 'Harry Potter' who smile back and actively listen to people.

References

- [1] N. Brooke and S. Scott. Computer graphics animations of talking faces based on stochastic models. In *International Symposium on Speech, Image Processing and Neural Networks*, 1994.
- [2] J. Cassel, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Ruloe-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *SIGGRAPH*, 1994.
- [3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.
- [4] O. Hasegawa, C.-W. Lee, W. Wongwarawipat, and M. Ishizuka. Realtime synthesis of moving human-like agent in response to user's moving image. In *Pattern recognition*, 1992.
- [5] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV*, 1996.
- [6] T. Jebara and A. Pentland. Action reaction learning : Automatic visual analysis and synthesis of interactive behaviour. In *ICVS'99*, pages 273–292, January 1999.
- [7] N. Johnson, , and D. Hogg. Learning the distribution of object trajectories for event recognition. In *Image and Vision Computing*, Aug. 1996.
- [8] N. Johnson. *Learning Object Behaviour Models*. PhD thesis, School of Computer Studies, University of Leeds, September 1998.
- [9] N. Johnson, A. Galata, and D. Hogg. The Acquisition and Use of Interaction Behaviour Models. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1998.
- [10] S. Morishima, Kiyoharu, and H. Harashima. An intelligent facial image coding driven by speech and phoneme. In *Acoustic, Speech and Sound processing*, 1989.
- [11] J. Olives, M. Sams, J. Kulju, and O. Seppala. Towards a high quality finnish talking head. In *Multimedia Signal Processing*, 1999.
- [12] H. Shatkey. The fourier transform - a primer. Technical report, Brown University, Department of Computer Science, 1995.
- [13] A. Takeuchi and K. Nagao. Communicative facial displays as a new conversational modality. Sony Computer Science Laboratory.