Model-based Synthesis of Visual Speech Movements from 3D Video

JAMES D. EDGE,
ADRIAN HILTON,
and
PHILIP JACKSON
The University of Surrey

In this paper we describe a method for the synthesis of visual speech movements using a hybrid unit selection/model-based approach. Speech lip movements are captured using a 3D stereo face capture system, and split up into phonetic units. A dynamic parameterisation of this data is constructed which maintains the relationship between lip shapes and velocities; within this parameterisation a model of how lips move is built and is used in the animation of visual speech movements from speech audio input. The mapping from audio parameters to lip movements is disambiguated by selecting only the most similar stored phonetic units to the target utterance during synthesis. By combining properties of model-based synthesis (e.g. HMMs, neural nets) with unit selection we improve the quality of our speech synthesis.

Categories and Subject Descriptors: J.0. [**Computer Applications**]: General

1. INTRODUCTION

Synthetic talking heads are becoming increasingly popular across a wide range of applications: from entertainment (e.g. Computer Games/TV/Films) through to natural user interfaces and speech therapy. This application of computer animation and speech technology is complicated by the expert nature of any potential viewer. Face-to-face interactions are the natural means of every day communication and thus it is very difficult to fool even a naïve subject that synthetic speech movements are real. This is particularly the case as the static realism of our models get closer to photo-realistic, whilst a viewer may accept a cartoon-like character readily they are often more sceptical of realistic avatars. To explain this phenomena Mori [1] posited the 'uncanny valley', the idea that the closer a simulcra comes to human-realistic the more slight discrepancies with observed reality disturb a viewer. Nevertheless, as the technology for capturing human likeness becomes more widely available the application of lifelike synthetic characters to the above mentioned applications has become attractive to our narcissistic desires. Recent films, such as the "The Curious Case of Benjamin Button", demonstrate what can be attained in terms of mapping captured facial performance onto a synthetic character. However, the construction of purely synthetic performance is a far more challenging task, and one which has yet to be fully accomplished.

The problem of visual speech synthesis can be thought of as the translation of a sequence of abstract phonetic commands into continuous movements of the visible vocal articulators (lips, jaw, tongue etc.) It is often considered that audible phonemes over specify the task for animation, that is an audio phoneme can discriminate based upon non-visible actions (e.g. voicing), thus visible-phonemes/visemes are often used as basis units for synthesis. The simplest attempts at synthesis often take static viseme units and interpolate between them in some manner to produce animation [2]. It should be noted that visemes in this context are often considered to be instantaneous static targets, whereas phonemes refer to a sequence of audio or vocal tract parameters. It is a limitation of this kind of approach that the dynamics of articulatory movement are often not included explicitly. In particular the context specificity of visemes must be modelled to correctly

synthesise speech, i.e. coarticulation. Viseme-interpolation techniques typically model coarticulation using a spline-based model to blend the specified targets over time [3], however, it is difficult to derive the parameters for such models from real articulatory data and it is not even known what shape the basis functions should take as they cannot be directly observed. Given these limitations current systems typically build models from the dynamics of the vocal tract which can be directly observed.

One of the most common techniques in audio speech synthesis is the selection and concatenation of stored phonetic units. By combining short sequences of real speech improvements in quality over parametric models of the vocal tract can be achieved. Analogously for visual synthesis short sections of captured speech movements can be blended together to produce animation. An example of this is Video-Rewrite [4] where short sections of video are blended together to produce what are termed video-realistic animations of speech. By indexing into real data unit-selection methods benefit from the intrinsic realism of the data itself. However, coarticulation is still manifest in how the units are blended together. It is not adequate to store a single unit for each phoneme; many examples must be stored across the various phonetic contexts and selected between during synthesis. In fact the best examples of concatenative synthesis select between speech units at different scales (e.g. phonemes, syllables, words etc.) to reduce the amount of blending and thus maximise the realism of the final animation. As the size of the underlying unit basis increases, the size of the required database exponentially increases, leading to a Catch-22 problem of database size vs. animation quality. Furthermore, concatenative techniques rarely take advantage of the audio dynamics when aligning units to the target utterance. It is necessarily true that the dynamics of articulatory movements are embedded within the audio itself, albeit perhaps sparsely, and this should be taken advantage of during synthesis.

The final group of visual synthesis techniques take advantage of the audio data to map into the space of visual speech movements. These audio-visual inversion models are typically based upon Hidden Markov Models (HMMs) [5], neural networks [6], or other lookup models [7]. Brand [5] constructed a HMM-based animation system to map from audio parameters (LPC/Rasta-PLP) to marker data which can be used to animate a facial model. The HMM is initially trained to recognise the training audio data, and for animation the output for each state is replaced by the distribution of visual parameters. Thus, a path through the hidden states of the HMM implies a trajectory through the articulatory space of a speaker. Problematically for this kind of model a HMM trained on audio data and another trained on the accompanying visual data would produce two very different network topologies. The approach of Brand makes the assumption that the two are at least similar, and this is unfortunately not the case. Constructing a global mapping in this way can produce a babbling level of synthesis, but does not accurately preserve the dynamics evident in the original training data. This can be improved by using HMMs representing smaller phonetic groupings (e.g. triphones), and using a lattice of these smaller units to both recognise the audio and animate the facial model. This is similar to the way that HMM speech recognition systems work; although in recognition we are making a binary decision, i.e. is this the correct triphone or not, whereas for animation we wish to recover a trajectory (sequence of states) that the vocal tract must pass through to produce the audio - a more difficult task. Also, because HMMs model speech according to the statistical mass of the training data the fine-scale dynamics of the individual trajectories can be lost in the mapping.

It can be seen that concatenative and model-based techniques have complementary features. In concatenative synthesis the fidelity of the original data is maintained, yet there is no global model of how lips move and a decision must be made on how to select and blend units. Model-based synthesis provides a global structure to constrain the movement of the articulators and traverses through this structure according to the audio of the target utterance, however, by matching the input audio to the statistical mass of training data the detailed articulatory dynamics can be lost. In this paper we use a hybrid approach which attempts to take the advantages of both models and combine them into a single combined system.

2. DATA CAPTURE

Many different forms of data has been used as the basis of visual speech synthesis. From photographs of visemes [11], frontal video of a speaker [2; 4], marker-based motion-capture data [10], and surface scans of a subject during articulation [12]. The research described in this paper is based on data recorded using a dynamic face capture system. This system works on the principal of stereophotogrammetry, where pairs of cameras are used to determine the location of points on a surface. The system consists of two stereo pairs (left/right) which use a projected infra-red pattern to aid stereo registration. Two further cameras capture colour texture information simultaneously with the surface geometry. All cameras operate at $60Hz$, and the output 3D models have in the order of $20,000$ vertices. Each frame of data is reconstructed independently, that is there is no initial temporal registration of the data. Audio data is captured simultaneously with the 3D geometry and texture.

To register the geometry over time markers are applied to the face of the subject. These take the form of blue painted dots on the skin and blue lipstick to track the contours of the lips. Between the markers alignment is performed by calculating the geodesic distance (i.e. across the surface of the skin) from a vertex in the first frame to its surrounding markers, in subsequent frames the location on the surface with the same relative position to surrounding markers is taken as the matching point. In this manner a dense registered surface reconstruction of the face can be captured for a subject. Due to the combination of the contour markers on the lips and the surface capture technology used we get a highly detailed model of the lips, in particular this is a great improvement over traditional motion-capture technology which is limited by the locations that markers can be attached to the face. We also get details of the movement of the skin surrounding the lips and in the cheeks which are commonly missed in synthesis systems. In the rest of this paper the data used is the registered 3D geometry, the texture images are only used to track the markers for registration. For the purposes of speech synthesis we isolate the data for the lower face (i.e. jaw, cheeks, lips) so that our system only drives the movement of the articulators.

The captured corpus consists of 8 minutes of registered 3D geometry and simultaneous audio captured of a male native British English speaker. Sentences were selected from the TIMIT corpus to provide a good sampling across all phonemes, there are 103 sentences in all (see table 2) and the sampling of phonemes can be seen in table 2. This does not represent a high sampling of phonemes in terms of context, as this was seen as too great a data capture effort to be feasible with the current equipment and time required to process the data. However, when considered as a reduced set of visemes, as opposed to phonemes, we have a relatively large set of exemplar animations in a high quality to facilitate the synthesis technique described in the following sections.

Herb's birthday occurs frequently on Thanksgiving.
She took it with her wherever she went.
Alice's ability to work without supervision is noteworthy.
Boy you're stirrin' early a sleepy voice said.
Employee layoffs coincided with companies reorganisation.
The armchair traveller preserves his illusions.
Don't ask me to carry an oily rag like that.
Why buy oil when you always use mine.
The sound of Jennifer's bugle scared the antelope.
Don't look for group valuables in a bank vault.
Continental drift is a geological theory.

Table I. Selected sentences from the corpus.

Consonants	p	72	b	79	m	99	ch	31
	jh	34	s	313	z	109	sh	41
	zh	20	f	69	v	58	th	28
	dh	81	k	133	g	39	t	241
	d	187	r	136	w	68	n	254
	ng	28	hh	29	l	170	y	62
Vowels	aa	24	ae	85	ah	48	ao	49
	aw	23	ay	57	ax	299	ea	26
	eh	73	ey	65	ia	22	ih	198
	iy	126	oh	62	ow	47	oy	24
	ua	23	uh	30				

Table II. Frequency of English phonemes in the captured data.

3. DATA REPRESENTATION AND CLUSTERING

The 3D registered data from the speech corpus is parameterised in a manner which facilitates the structuring of a state-based model. The dataset consists of a sequence of frames, F, where the i^{th} frame $F_i = \{x_0, y_0, z_0, \ldots, x_i, y_i, z_i, \ldots, x_n, y_n, z_n\}$. Principal Component Analysis (PCA) is applied directly to F to filter out low variance modes. By applying PCA we get a set of basis vectors, $\overrightarrow{X}$. The EM method for computing principal components [9] is used here due to the size of the data matrix, F, which holds $28,833$ frames $\times 12,784$ xyz coordinates. The first 100 basis vectors are computed, with the first 30 holding over 99% of the recovered variance. The percentage of the total variance accounted for will be lower, but the scree-graph shows that the important features of F are compressed in only a few dominant components (i.e. $\sim 95\%$ in the first 10 components, and $\sim 99\%$ in the first 30 components indicating a flattening of the scree-graph.) F can be projected onto the basis $\overrightarrow{X}$ to produce the parameterisation F^x. So each frame F_i can be projected onto $\overrightarrow{X}$, $F_i \times \overrightarrow{X} \rightarrow F_i^x$. Broadly, the 1^{st} component of $\overrightarrow{X}$ can be categorised as jaw opening, and the 2^{nd} is lip rounding/protrusion, lower variance components are not as easily contextualised in terms of observed lip-shape qualities but generally describe protrusion, asymmetries and the bulging of the cheeks.

The first derivative for each frame can be estimated as $F_i^{x\prime} = F_i^x - F_{i-1}^x$ (the parametric displacement of the lips in $1/60^{th}$ of a frame.) Each pair $\{F_i^x, F_i^{x\prime}\}$ describes a distinct state in the physical space of lip movement. Another level of PCA could be applied directly upon this state data, however as the first derivative is at a different scale the parameters need to be normalized such that F_i^x does not dominate over $F_i^{x\prime}$. Thus a state matrix $S = \{\alpha(F_i^x - \mu), \beta(F_i^{x\prime} - \mu')\}$ is constructed where all parameters are scaled to the range $[-1, 1]$. This state matrix is now processed in a manner similar to Multidimensional Scaling (MDS), that is a symmetric distance matrix D is formed where each element D_{ij} is the euclidean distance between the states S_i and S_j, i.e. $D_{ij} = \sqrt{(S_i - S_j)^2}$. The matrix D is then decomposed using another iteration of PCA forming a basis $\overrightarrow{Y}$, so for each of the initial frames F_i we have a corresponding projection into the state space F_i^y. The first 3 dimensions of $\overrightarrow{Y}$ account for over 93% of of the recovered variance in D.

The described parameterisation is used to reduce the dimensionality from $\sim 38,000$ dimensions down to 15 dimensions. The manifold evident in this reduced space also demonstrates several properties that are of interest for the visualisation of articulatory dynamics, in particular with regards the cyclical nature of speech lip movements which are evidenced in the symmetric nature of the manifold. A discussion of the properties of the speech manifold can be found in [8]. As this parameterisation maintains the relationship between lip shapes and their derivatives it is ideal for structuring a state-based model of speech movements. For the purposes of speech synthesis we use the reduced space to cluster the data, where each individual cluster represents a dynamic state in the system. Clustering is performed in this manner to avoid the dimensionality problem which would make clustering of the raw data computationally expensive and error prone. Furthermore by clustering according to both position and velocity we implicitly pre-structure our state-based model of speech articulation

4. SYNTHESIS OF SPEECH LIP MOVEMENTS

Synthesis of speech lip movements in our system is characterised by a hybrid approach that combines unit selection with a model-based approach for traversing the space of the selected phonemes. This can be seen as a traversal of a subspace on the manifold of lip motion described in the previous section. By cutting down the possible paths, according to the input audio, we reduce the ambiguity of the mapping from audio to visual speech movements and produce more realistic motions. The input to our system is a combination of both a phonetic transcription and the audio for the target utterance. Some systems attempt to avoid the necessity for a phonetic transcription by using a model which is effectively both recognising the phonetic content and synthesising the visual component simultaneously, or which forego any phonetic structure and

attempt to directly map from audio parameters to the space of visual movements [5; 7]. In our experience recognition and synthesis are very different problems and improved results can be attained by seperating the recognition/transcription component, which can be dealt with either using a specialised recognition module or manually depending upon the requirements of the target application.

Synthesis proceeds by taking the phonetic transcription and the audio for the target utterance (decomposed into MFCCs) and selecting for each segment the most similar stored phonetic exemplar. A phoneme for our purposes consists of the sequence from the centre of the preceding phoneme to the centre of the following phoneme, similar to a triphone but only classified according to the central phonetic content (i.e. not according to context.) The distance between a segment of the target utterance and a phonetic exemplar is calculated using Dynamic Time Warping (DTW.) This algorithm calculates the minimum aligned distance between two time-signals using a recursive algorithm (1).

$$d_{i,j} = \sqrt{(x_i - y_j)^2}$$

$$D_{i,j} = min \left\{ \begin{array}{c} D_{i-1,j} + d_{i,j} \\ D_{i,j-1} + d_{i,j} \\ D_{i-1,j-1} + 2d_{i,j} \end{array} \right\} \quad (1)$$

Here $d_{i,j}$ is the local distance between two frames of input data x_i and y_j, and $D_{i,j}$ is the global distance accumulated between the sequences $x \in [1, i]$ and $y \in [1, j]$. The smallest global matching distance between the segment from the target utterance and an exemplar from the stored dataset indicates the best available unit. Note that this does not require the input transcription to be fully accurate as the algorithm will find the best alignment between the two sequences to calculate the global distance between phonetic units.

Usually in unit selection synthesis models the motions are blended directly to produce a contiguous animation trajectory. This is problematic as the boundaries of the units may not align well leading to jumps in the animation. However, if the units are selected to allow good transitions then they may not be optimal for the target utterance. Furthermore, some phonemes have a stronger effect upon the output motion than others and it would be advantageous to use the evidence available in the target audio to determine the final trajectory. In our system we select the best units given the target audio, as described, and use a model-based approach built from these units to determine a global trajectory for the target utterance.

A state-based model is built to model the global dynamics of speech lip motion. States are clusters forming a discretisation of the speech manifold described in Section 3. The model we use consists of $N = 200$ states each of which corresponds to a single distribution of lip shapes/velocities. An $N \times N$ transition matrix, T, is also constructed with each element $T_{i,j}$ containing 0 to indicate connected states and ∞ to indicate unconnected states. Given that states are clustered on both position and velocity, the transition matrix is an implicit constraint upon the second derivative (acceleration) of speech lip movements. Note that this model is entirely built on the space of visual movements; i.e. this is the opposite to models such as [5] where the state-based model is initially trained on the audio data. Each of these states will correspond to a range of possible audio parameters. In fact the range of possible audio parameters that correspond to a single dynamic state can be widely distributed across the space of all speech audio. This is problematic for a probabilistic HMM approach which attempts to model these distributions using Gaussian Mixture Models (GMMs.) Instead we consider each example within a state to be independent rather than a part of a probabilistic distribution, and use the best available evidence of being in a state to traverse the model and generate a synthetic trajectory. The structure of the state model is constructed as a pre-processing step using the entire dataset.

To generate a trajectory from the state-based model we use a *Viterbi*-like approach, albeit to calculate a path using a minimum aligned distance criteria and not a maximum probability. The algorithm proceeds

by calculating a state distance matrix S^d of size $N \times L$ (i.e. number of states $\times$ number of frames in the target utterance.) Each element $S^d_{i,j}$ contains the minimum distance between the i^{th} frame of input data to all the contextually relevant frames in state j. The distance between a frame of audio data and a state will change according to its phonetic context in the target utterance, and in this way we optimize the mapping from audio $\rightarrow$ visual parameters according to the selected units. If we have a sequence of P phonemes this is similar to training $P-1$ models, one for each phoneme-phoneme transition in the sequence, during synthesis (i.e. not as a pre-processing step.)

Each element of S^d, $S^d_{i,j}$, is a minimum distance value between a window surrounding the i^{th} frame of audio data from the target utterance and each of the contextually relevant examples in S^d. We use a window size of 5 frames to perform this distance calculation, multiplied by a *Gaussian* windowing function, α, to emphasise the importance of the central frame. The distance function, *dist*, between an input window of audio data, u, and a state in the context of its left and right selected units, S^d_{lr}, is defined in (2), where each v is an independent example from S^d_{lr}.

$$\begin{aligned} u &= \{\alpha(-2)x_{t-2}, \ldots, \alpha(0)x_t, \ldots, \alpha(2)x_{t+2}\} \\ v &= \{\alpha(-2)y_{s-2}, \ldots, \alpha(0)y_s, \ldots, \alpha(2)y_{s+2}\} \\ dist(u, S^d_{lr}) &= min\{sqrt(u-v)^2\}, \forall v \in S^d_{lr} \end{aligned} \tag{2}$$

To calculate the optimal trajectory across the speech manifold, we perform a simple recursive algorithm to accumulate distance according to the allowable transitions in T. The accumulated distance matrix, S^D, is calculated according to the recursion in (3)[1].

$$S^D_{i,j} = min\{S^D_{i-1,k} + T_{k,j} + S^d_{i,j}\}, k \in [1, N] \tag{3}$$

This is a simple distance accumulation operation with the transition matrix ensuring that states can only be jumped between if that transition was seen in the original dataset. The minimum distance to a state at frame L identifies the optimal alignment, and by maintaining back-pointers the sequence of states can be traced back through S^D. The output at this stage of synthesis is a sequence of states, where each state is characterised by a distribution of visual parameters.

5. ANIMATION

Each frame of output from the synthesis procedure outlined in the previous section is a 3D surface scan of the same form tracked in the original data. This means that we only have surface detail for the region of the face bounded by the tracked markers. Because markers cannot be placed in regions of shadow or where occlusions may occur we do not have geometry for the region between the neckline and the jaw. Also, as the colour texture from the dynamic scanner contains markers it is impractical to use for display. For these reasons we need to supplement the data originally captured to produce a photo-realistic rendered animation.

Jaw rotation is modelled using a morph-target model. Scans from a static surface scanner are used to model a 1D jaw rotation parameter, i.e. in-between shapes are taken as an alpha-blend between the two extrema. Whilst generally this is inadequate, as the jaw has more than a single degree-of-freedom, for speech the majority of jaw movement can be characterised as an opening/rotation action. The jaw rotation parameter for each frame in a synthetic sequence can be simply derived using a 1D line search optimisation which minimises the distance between the jaw model and the synthetic mouth data. The lower teeth are also attached to the jaw model so that we have detail within the mouth.

[1]This recursion is virtually identical to the *Viterbi* algorithm (when using log probabilities), the difference being that *Viterbi* is probabilistic whereas here we are simply accumulating distances and only use a binary transition matrix.

The results shown in this paper are produced by warping a single image using the synthetic mouth data and the fitted jaw model. This is done using a layered model where the image is progressively warped at each level to produce each output frame. The optimal projection of the jaw model into the image plane is calculated along with the non-rigid alignment with facial features in the photograph, using this information the image can be warped to fit the required jaw rotation. The synthetic mouth data is simply overlayed on top of the jaw animation using a second image warping operation. Because the image itself is not parameterised, as in active appearance models [], we maintain the quality of the image itself after animation. Furthermore, because a true 3D model underlies the synthesis the same technique could be potentially used on video sequences with extreme changes in head pose, which is generally problematic for purely 2D methods [2; 4].

6. EVALUATION

7. CONCLUSIONS

REFERENCES

[1] M. Mori. *The Uncanny Valley* (K.F. MacDorman & T. Minato, Trans.). Energy, 7(4), 33–35, 1970.

[2] T. Ezzat, G. Geiger, and T. Poggio. *Trainable Videorealistic Speech Animation*. Proceedings of SIGGRAPH'02, ??–??, 2002.

[3] M.M. Cohen and D.W. Massaro. *Modeling coarticulation in synthetic visual speech*. Models and Techniques in Computer Animation, 1993.

[4] C. Bregler, M. Covell and M. Slaney. *Video Rewrite: driving visual speech with audio.*. Proceedings of SIGGRAPH'97, 1997.

[5] M. Brand. *Voice puppetry*. Proceedings of SIGGRAPH'99, 1999.

[6] D.W. Massaro, J. Beskow, M.M. Cohen, C.L. Fry, and T. Rodriguez. *Picture my voice: Audio to visual speech synthesis using artificial neural networks*. Proceedings of AVSP'99: International Conference on Auditory-Visual Speech Processing, 133–138, 1999.

[7] B. Theobald, and N. Wilkinson. *A probabilistic trajectory synthesis system for synthesising visual speech*. Proceedings of Interspeech'08, 2008.

[8] J.D. Edge, A. Hilton, and P. Jackson. *Parameterisation of 3D Speech Lip Movements*. Proceedings of AVSP'08: International Conference on Auditory-Visual Speech Processing, 2008.

[9] S. Roweis. *EM Algorithms for PCA and SPCA*. Proceedings of NIPS'97, 626–632, 1997.

[10] Y. Cao, P. Faloutsos, and F. Pighin. *Expressive Speech-Driven Facial Animation*. ACM Transactions on Graphics, 24(4), 1283–1302, 2005.

[11] T. Ezzat, and T. Poggio. *Videorealistic Talking Faces: A Morphing Approach*. Proceedings of AVSP'97, 1997.

[12] P. Mueller, G. A. Kalberer, M. Proesmans, and L. Van Gool. *Realistic Speech Animation Based on Observed 3D Face Dynamics*. IEE Proc. Vision, Image & Signal Processing, 152, 491–500, 2005.