Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 597267, 12 pages
doi:10.1155/2009/597267

Research Article

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

James D. Edge, Adrian Hilton, and Philip Jackson

Centre for Vision, Speech and Signal Processing, The University of Surrey, Surrey GU2 7XH, UK

Correspondence should be addressed to James D. Edge, j.edge@surrey.ac.uk

Received 1 March 2009; Revised 30 July 2009; Accepted 23 September 2009

Recommended by Gérard Bailly

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.

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 photorealistic. 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 (e.g., lips, jaw, tongue). It is often considered that audible phonemes overspecify the task for animation, that is, an audio phoneme can discriminate based upon nonvisible actions (e.g., voicing in pat versus bat), and thus visible-phonemes/visemes (a term coined by Fisher [2]) 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 [3–6]. 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 kinematics of articulatory movement are often not included explicitly. In particular the context specificity of visemes must be modelled to correctly synthesise speech, that is, coarticulation. Viseme-interpolation techniques typically model coarticulation using a spline-based model (with reference to Löfqvist's earlier work on coarticulation [7])

to blend the specified targets over time [6]. 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 kinematics of the vocal tract which can be directly observed. In [8] motion-captured markers (Optotrak) are recorded for natural speech for a single speaker; these are then used to train the parameters for an adapted version of the authors' earlier coarticulation model [6]. In [5] tracked markers of isolated French vowels and VCV syllables are used to train the parameters from Öhman's numerical model of coarticulation [9]. In [3] video of a speaker is used to train the distribution of visual parameters for each viseme, with synthesis performed by generating a trajectory that passes through the relevant distributions. In [10] viseme transition functions for diphones and triphones are trained using motion capture data, combinations of which can be used to synthesise novel utterances.

One of the most common techniques in audio speech synthesis is the selection and concatenation of stored phonetic units (e.g., Festival [11], MBROLA [12]). 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 [13] where short sections of video are blended together to produce what are termed video-realistic animations of speech. In [14, 15] motion-captured marker data is concatenated to similar effect, albeit without the advantage of photorealistic texture. Cao et al. [16] use similarity in the audio parameters between stored units and the target utterance as a selection criterion, along with terms which minimize the number of units and cost of joining selected units. 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) to reduce the amount of blending and thus maximise the realism of the final animation (this is effectively being done in [16]). As the size of the underlying unit basis increases, the size of the required database exponentially increases; this leads to a trade-off between database size and animation quality.

The approaches described thus far do not use the audio of the target utterance to guide the generation of a synthetic speech trajectory. It is necessarily true that 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) [17, 18], neural networks [19], or other lookup models [20]. Brand [18] constructed an 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 audio data, and for animation the output for each state is replaced by the corresponding distribution of visual parameters. Thus, a path through the hidden states of the HMM implies a trajectory through the articulatory space of a speaker. Zhang and Renals [17] use a trajectory formulation of HMM synthesis to synthesise Electro-Magnetic Articulography (EMA) trajectories from the MOCHA-TIMIT corpus. Trajectory HMMs incorporate temporal information in the model formulation which means that they generate continuous trajectories and not a discrete sequence of states. Problematically for all HMM synthesis a model 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 motion 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, that is, 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 structure of the individual trajectories can be lost in such a mapping.

In order to capture speech articulatory movements several methods have been used; these include photography/video [3, 13, 21], marker-based motion capture [8, 10, 14, 15], and surface-capture techniques [22–25]. Video has the advantage of realism, but because the view is fixed, the parameters of such models do not fully capture the variability in human faces (e.g., in the absence of depth, lip protrusion is lost). Marker-based motion capture systems allow the capture of a small number of markers (usually less than 100) on the face and provide full 3D data. However, marker-based systems are limited by the locations in which markers can be placed; in particular the inner lip boundary cannot be tracked which is problematic for speech synthesis. Furthermore, systems such as Vicon and Optotrak require the placement of physical markers and sometimes wires on the face which do not aid the subject in speaking in a natural manner. Surface capture technologies, usually based upon stereophotogrammetry, produce sequences of dense scans of a subject's face. These are generally of a much higher resolution than possible with marker-based mocap (i.e., in the order of thousands of vertices), but frames are generally captured without matching geometry over time. This unregistered data requires a second stage of alignment before it can be used as an analytical tool.

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 movements 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. The most similar approach to that described can be found in [26] where an HMM model is used together with a concatenation approach for speech synthesis of both audio and visual parameters. However, Govokhina et al. use a HMM to select units for concatenation, whereas we select units to train a state-based model for synthesis (i.e., effectively the opposite order). The data used comes from a high-resolution surface capture system combined with marker capture to aid the registration of face movements over time. This paper is structured in the following manner: Section 2 describes our dynamic face capture and the makeup of our speech corpus; Section 3 describes the parameterisation of this data and the recovery of an underlying speech behaviour manifold; Section 4 describes our approach to the synthesis of speech lip movements; Section 5 describes the rendering/display of synthetic speech animation on a photorealistic model; finally, Section 6 discusses a perceptual evaluation study into the quality of our synthesis approach.

2. Data Capture

Many different forms of data have been used as the basis of visual speech synthesis: from photographs of visemes [21], frontal video of a speaker [3, 13], marker-based motion-capture data [16], and surface scans of a subject during articulation [23]. The research described in this paper is based on data recorded using the 4D capture system developed by 3dMD [27] for high-resolution capture of facial movement; see Figure 1(a). 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 have a resolution of 1.2 Megapixels and operate at 60 Hz, and the output 3D models have in the order of 20 000 vertices (full face ear-to-ear capture). Each frame of data is reconstructed independently; this means that there is no initial temporal registration of the data. Audio data is also 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; see Figure 1(b). 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. During data capture the subject is asked to keep their head still to prevent them leaving the capture volume which is relatively restrictive. However, no physical constraint is applied and it is found that the subject's head will drift slightly during recording (a maximum 2 minutes of continuous data capture is performed) which is removed using the Iterative Closest Point (ICP [28]) rigid alignment algorithm.

Table 1: Selected sentences from the corpus.

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 are 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

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 [29] to provide a good sampling across all phonemes, there are 103 sentences in all (see Table 1, e.g., sentences), 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. The audio data is manually transcribed to allow both the audio and geometry data to be cut into Phone segments.

(a) (b)

Figure 1: Capture of facial movements: (a) the face capture system; (b) frames and tracked geometry from a sequence in the captured dataset.

Table 2: Frequency of English phonemes in the captured data.

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				

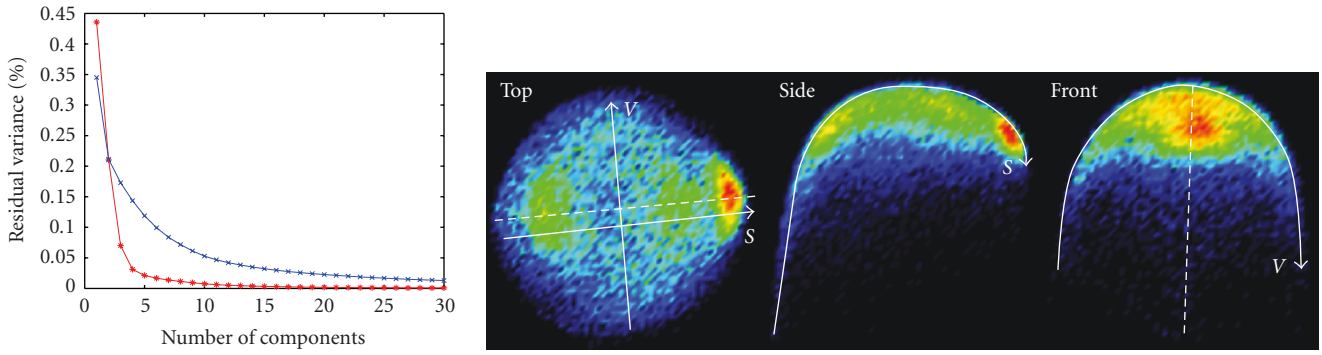
3. Data Parameterisation

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 ith frame $F_i = \{x\vec{y}z_0, x\vec{y}z_1, \ldots, x\vec{y}z_i, \ldots, x\vec{y}z_n\}$ and $x\vec{y}z$ is a 3D vertex. 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, $\vec{X}$. The EM method for computing principal components [30] 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., ~95% in the first 10 components and ~99% in the first 30 components indicating a flattening of the scree-graph, see the blue line in Figure 2(a)). F can be projected onto the basis $\vec{X}$ to produce the parameterisation F^X. So each frame F_i can be projected onto $\vec{X}$, $F_i \times \vec{X} \rightarrow F_i^X$. Broadly, the 1st component of $\vec{X}$ can be categorised as jaw opening, the 2nd is lip rounding/protrusion, and 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 $F_i^{X'} = F_i^X - F_{i-1}^X$ (the parametric displacement of the lips in 1/60th of a second). Each pair $\{F_i^X, F_i^{X'}\}$ describes a distinct point in the physical space of lip movement. Another level of PCA could be applied directly upon this 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 $F_i^{X'}$. Thus a matrix $M = \{(1/\sigma^2)(F_i^X - \mu), (1/\sigma'^2)(F_i^{X'} - \mu')\}$ is constructed where the F_i^X and $F_i^{X'}$ are scaled to have unit variance.

The matrix M is now processed in a manner similar to Multidimensional Scaling (MDS) [31]; that is, a symmetric distance matrix Δ is formed where each element Δ_{ij} is the Euclidean distance between M_i and M_j (the ith and jth elements of M), that is, $\Delta_{ij} = \sqrt{(M_i - M_j)^2}$. The matrix Δ is then decomposed using another iteration of PCA forming a basis $\vec{Y}$; so for each of the initial frames F_i we have a corresponding projected coordinate F_i^Y. The first 3 dimensions of $\vec{Y}$ account for over 93% of the recovered variance in Δ.

The described parameterisation is used to reduce the dimensionality from 38,352 (number of vertices $\times$3) dimensions down to 10 dimensions, which account for ~99%



(a) (b)

FIGURE 2: Parameterisation of speech lip movements: (a) residual variances for the first 30 dimensions of $\vec{X}$ (blue) and $\vec{Y}$ (red); (b) the speech manifold evident in the first 3 dimensions of $\vec{Y}$: colour indicates density of the projection (blue least dense → red most dense), dashed line indicates the plane of symmetry between opening/closing of the lips, and vectors S and V indicate maximum change in lip shape and velocity, respectively.

of the variance in Δ (as shown in the scree plot, see the red line in Figure 2(a)). The manifold evident in this reduced space also demonstrates several properties that are of interest for the visualisation of articulatory movements. The first 3 dimensions of the recovered speech manifold are shown in Figure 2(b). The major properties of this manifold are an ordering of frames according to change in both lip shape (the non linear vector S) and velocity (the nonlinear vector V). The manifold is also symmetric about a plane which divides lip-opening states from lip-closing states, and as a consequence of this speech trajectories are realised as elliptical paths on the manifold (i.e., open-close-open cycles). This structured representation is useful for the visualisation of speech movements, and a more detailed discussion of the properties of the recovered speech manifold can be found in [22]. 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, Y, to cluster the data, where each individual cluster represents a state of motion 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 prestructure our state-based model of speech articulation discussed in the next section. Details of the state clustering and model construction are given in Section 4.

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 synthetic 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 that 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 [18, 20]. In our experience, recognition and synthesis are very different problems and improved results can be attained by separating the recognition and transcription component, which can be dealt with either using a specialised recognition module or manually depending upon the requirements of the target application.

In overview, see Figure 3, our system proceeds through the following steps.

(1) Input audio is decomposed into Mel Frequency Cepstral Coefficients [32] (MFCCs), and a phonetic transcription of the content.

(2) A unit selection algorithm is used to determine the closest stored unit to each segment in the target utterance.

(3) Selected units are used to train a state-based model for each phone-phone transition.

(4) An optimal path through the trained model, that is, across the learned manifold from Section 3, is determined using a *Viterbi* type algorithm.

(5) The recovered sequence of states, which map onto a sequence of distributions of lip shapes/velocities, is used to generate a smooth output trajectory for animation.

Synthesis begins by taking the phonetic transcription and the audio for the target utterance (decomposed into 12th order MFCCs at the same frame rate as the geometry,

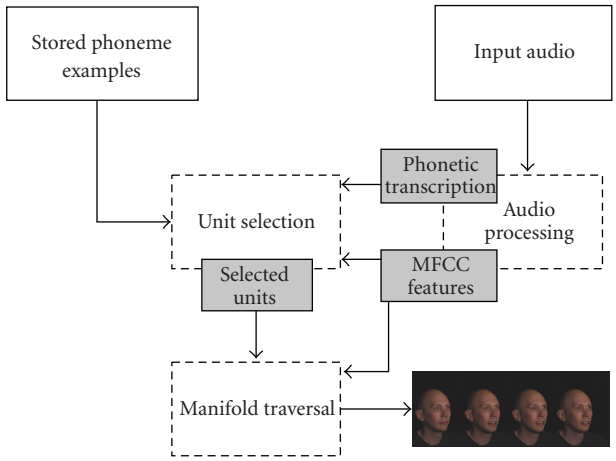


Figure 3: Schematic of the synthesis process: stored phoneme exemplars along with the input audio features are used to select optimal units to train a state-based manifold traversal model.

60 Hz) and selecting for each segment the most similar stored phone. A phone for our purposes consists of the sequence from the centre of the preceding phone to the centre of the following phone, similar to a triphone but only classified according to the central phone (i.e., not according to context). The distance between a segment of the target utterance and a stored phone is calculated using Dynamic Time Warping (DTW). This algorithm calculates the minimum aligned distance between two time-series using using the following recursive equation:

$$d_{i,j} = \sqrt{\left(x_i - y_j\right)^2},$$
$$D_{i,j} = \min\begin{Bmatrix} D_{i-1,j} + d_{i,j} \\ D_{i,j-1} + d_{i,j} \\ D_{i-1,j-1} + 2d_{i,j} \end{Bmatrix}. \quad (1)$$

Here $d_{i,j}$ is the local Euclidean distance between a frame of the input data x_i and a frame from a stored exemplar 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 because the algorithm finds the best alignment between the two sequences, small inaccuracies in the input transcription will not reduce the quality of the final animation. This is in contrast to other concatenative synthesis systems (e.g., [13, 15]) where the accuracy of the transcription is key to producing good results. Our system aligns to the audio itself rather than to a, potentially inaccurate, transcription.

Usually in unit selection synthesis models, the motions are blended directly to produce a continuous 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 above, 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 fit the input audio to the global structure of speech lip movements stored in our dataset. States are clusters forming a discretisation of the speech manifold described in Section 3. We use the bisecting K-means algorithm to cluster the parameterised data into states. The model we use consists of $N = 200$ states, each of which corresponds to a single distribution of lip shapes and velocities. The number of states is chosen as a trade-off between dynamic fidelity (i.e., a higher number of states gives a more accurate representation of speech movements), database size (i.e., the number of states must be much less than the number of samples in the dataset), and processing time (i.e., more states take longer to produce a global alignment). An $N \times N$ binary transition matrix, T, is also constructed with each element $T_{i,j}$ containing 0 to indicate connected states and ∞ to indicate unconnected states. A connection in $T_{i,j}$ means that a frame from the captured dataset classified in state i is followed by a frame classified in state j. 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; that is, this is the opposite to models such as [18] where the state-based model is initially trained on the audio data. Each of our 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 that models these distributions using Gaussian Mixture Models (GMMs) and has an underlying assumption that they are relatively well clustered. 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 choice of using a binary transition matrix (i.e., not probabilistic as in a HMM) also means that transitions which occur infrequently in the original data are equally as likely to be traversed during synthesis as those which are common. In this way we increase the importance of infrequent sequences, maximising the use of the captured data. The structure of the state model is constructed as a preprocessing step using the entire dataset.

To generate a trajectory from the state-based model we use a dynamic programming approach similar to *Viterbi*, albeit to calculate a path using a minimum aligned distance criteria and not maximum probability. The algorithm proceeds by calculating a state distance matrix S^d of size $L \times N$ (i.e., number of frames in the target utterance $\times$ number of states). Each element $S^d_{i,j}$ contains the minimum Euclidean cepstral distance between the ith frame of input data to all the contextually relevant frames in state j. Here

a frame from state j is considered only if it is from one of the previously selected units which bracket frame i (i.e., the selected left-right phonetic context of the frame). Because of this the distance between a frame of audio data and a state will change according to its phonetic context in the target utterance. This optimises the mapping from audio to 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 preprocessing step).

Each element of S^d, $S^d_{i,j}$, is a minimum distance value between a window surrounding the ith frame of audio data from the target utterance and each of the contextually relevant examples in state S_j. We use a window size of 5 frames to perform this distance calculation, multiplied by a *Gaussian* windowing function, $\gamma(n) = (1/\sqrt{2\pi})\exp^{(-n^2/2)}$, to emphasise the importance of the central frame. The distance function, dist, between an input window of audio data, u, at time i, and a state in the context of its left and right selected units, S^{lr}_j, is defined in (2) where each v is a window of audio frames, centred at time k, from either the left or right selected units at this point in the sequence (i.e., where $v \in S^{lr}_j$). The x and y are individual frame samples from each of the windows, u and v, respectively,

$$\begin{aligned} u_i &= \{\gamma(-2)x_{i-2}, \ldots, \gamma(0)x_i, \ldots, \gamma(2)x_{i+2}\}, \\ v_k &= \{\gamma(-2)y_{k-2}, \ldots, \gamma(0)y_k, \ldots, \gamma(2)y_{k+2}\}, \\ S^d_{ij} &= \operatorname{dist}\left(u_i, S^{lr}_j\right) = \min\left\{\sqrt{(u_i - v_k)^2}\right\}, \quad \forall v_k \in S^{lr}_j. \end{aligned} \tag{2}$$

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 the following equation:

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

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. Equation (3) is a simple distance accumulation operation with the transition matrix ensuring that transitions between states can only occur if that transition was seen in the original dataset. The minimum distance to a state at frame L identifies the optimal alignment. By maintaining back-pointers the sequence of states can be traced back through S^D.

One problem with the proposed method is that by only selecting the best units for training the state-based model, there is a possibility that the model cannot transition between two neighbouring selected units. This could occur, for example, if the context for the selected units means that the boundaries are very far apart. Constraints on the size of database we can capture means that it is impossible to store exemplars for all phonemes in all contexts. Thus a back-off solution for this problem is used. The point at which the model has failed to transition is simple to find, given that S^D will contain ∞ for all columns past this point. We can add examples from the dataset, in order of similarity to the target audio which will weaken the initial constraint on which parts of the speech manifold can be traversed. This is done by selecting the next most similar unit for the left and right context at this point in the sequence and adding the frames from these examples to each of the S^{lr} context states. So the S^{lr} are initially trained on the two most similar phones for the context, then four, then six, and so forth until the algorithm can pass through the segment. In practice, this is an infrequent problem and this solution does not add greatly to the complexity of the algorithm (given that we have already calculated a ranking of similarity between each input segment and all relevant stored examples).

The output at this stage of synthesis is a sequence of states, where each state is characterised by a distribution of visual parameters. Given that for each state we have a distribution of positions and velocities for the lips, we use Brand's [18] approach for deriving a continuous trajectory. Each state has a mean position μ_i and velocity μ'_i as well as a full-rank covariance matrix C_i relating positions and velocities. For a sequence of states, $S = \{S_1, \ldots, S_i, \ldots, S_L\}$, and frame parameters $Z = \{z_1, \ldots, z_i, \ldots, z_L\}^T$ (where z_i is a vector containing both the position and velocity at time i) this can be formulated as a maximum likelihood problem:

$$Z^* = \arg\max_Z \log \prod_i \mathcal{N}\left(\tilde{z}_i; C_{S(i)}\right). \tag{4}$$

In (4) $\mathcal{N}(z; C)$ is the Gaussian probability of $\tilde{z}$ according to the state covariance matrix C where $\tilde{z}$ is mean centered. The optimal trajectory, Z^*, of this formulation can be found by solving a block-banded system of linear equations. The output is a continuous trajectory of parameters, which yields a smooth animation of lower facial movement of the same form seen in our database (see Figure 6 for examples of the output 3D meshes from synthesis). Processing time for the sentences from our dataset, including both model building and synthesis, was in the range 30–50 seconds, depending upon the length of the target utterance. Figure 4 shows several examples of synthesised trajectories next to the real data for utterances in the dataset (the sentences were held out of the training set for synthesis). Section 5 discusses how this is turned into a photoreal animation of a speaker for display.

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 (i.e., geometry of the lower face). 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 photorealistic rendered animation. Note that the synthesis results from the previous

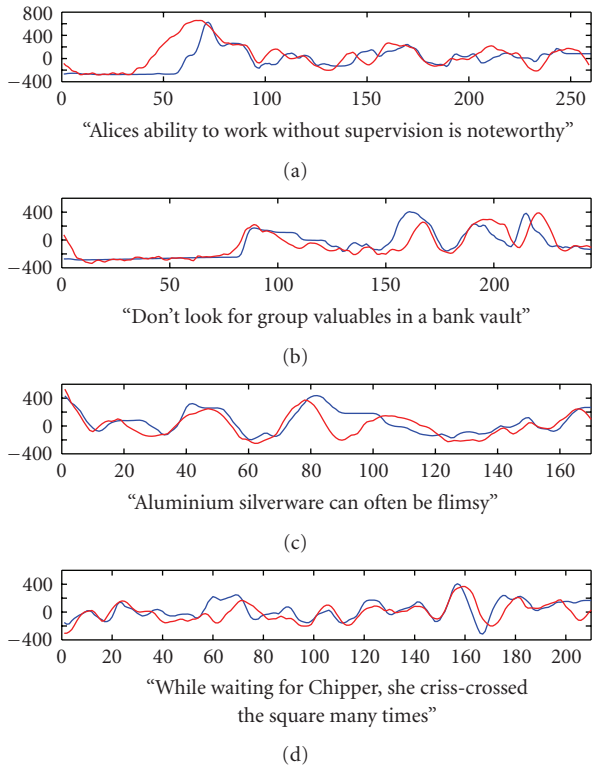


Figure 4: Comparison of synthesised trajectories using our approach (blue) and real data (red) for the first dimension of the PCA model $\vec{X}$.

section are used to animate the lower face, and the following model is used only to integrate this into a full face model.

In the animation results, jaw rotation is modelled using a 3D morph-target model. Scans from a static surface scanner are used to model a 1D jaw rotation parameter; that is, in-between shapes are taken as an alpha-blend between two extrema (shown in Figure 5). Generally this is inadequate, in [33, 34] the 6 degrees-of-freedom of the jaw are examined in detail, but for our purposes where only speech movements of relatively low amplitude are being synthesised a single degree-of-freedom has been found to be adequate (i.e., the join between the synthesis results and the jaw model is not noticeable). It is important to note that the original captured data includes the actual motion of the jaw, and this 1D model is only intended to fill in the region beneath the jawline to prevent a discontinuity in the rendered results. The jaw model is fitted to the synthesis results by performing a 1D line search to find the position at which the jawline of the synthetic lower face geometry fits that of the jaw model. The function, $f(\alpha)$, which defines the goodness of fit of the jaw model given a particular interpolation parameter, α, is shown in the following equation:

$$f(\alpha) = \sum_i s_i - \left(\alpha \cdot t_i^0 + (1-\alpha)\cdot t_i^1\right), \quad \alpha \in [0,1]. \tag{5}$$

In this equation the s_i are the jawline vertices for a frame of the synthesised lower face geometry, and the t_i^0 and t_i^1 are the matching vertices of the jaw model for the two extrema

Figure 5: Jaw rotation morph targets.

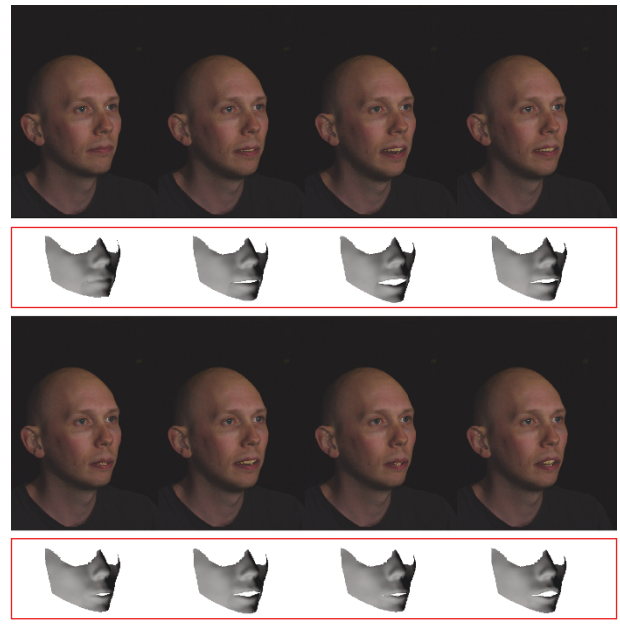


Figure 6: Rendered frames and generated 3D meshes (in red boxes) for the utterance "*Morphophonemic rules may be thought of as joining certain points in a system*".

(closed and open, resp.). Newton's method with derivatives calculated by finite differences is used to find the minima of (5), which is adequate as there is only a single minima within the range $\alpha \in [0,1]$. For the purposes of fitting the jaw model it is important that the jaw extrema are chosen such that they bracket the range of speech movements during normal speech.

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 nonrigid 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. This is similar to the work of [35], albeit our model is purely 3D. Because the image itself is not parameterised, as in

Table 3: The mean and variance of responses for the naturalness evaluation study; the three cases are real data playback ($\mu_{\text{real}}, \sigma^2_{\text{real}}$), synthetic trajectories using the technique described in this paper ($\mu_{\text{synth}}, \sigma^2_{\text{synth}}$), and synthetic trajectories using viseme interpolation ($\mu_{\text{interp}}, \sigma^2_{\text{interp}}$).

Subject	μ_{real}	σ^2_{real}	μ_{synth}	σ^2_{synth}	μ_{interp}	σ^2_{interp}
Subject 1	3.45	1.11	2.95	1.37	2.63	1.95
Subject 2	4.00	0.85	3.22	1.13	2.14	0.69
Subject 3	3.55	0.74	2.90	1.51	1.73	0.87
Subject 4	3.84	0.62	3.11	0.85	3.07	0.44
Subject 5	3.32	0.79	2.73	0.39	2.27	0.68
Subject 6	3.68	1.17	3.55	0.92	2.64	0.81
Subject 7	4.36	0.43	3.90	0.65	3.13	0.59
Overall	3.74	0.89	3.19	1.09	2.52	1.05

active appearance models [36], we maintain the quality of the image itself after animation (i.e., we do not get the blurring associated with such models). 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 (such as [3, 13]). Frames from a synthetic sequence for the sentence "*Morphophonemic rules may be thought of as joining certain points in a system*" are shown in Figure 6.

The major problems in the animation of our model are the missing features, in particular the lack of any tongue model. Ideally we would also animate the articulation of the tongue; however, gathering dynamic data regarding tongue movement is complex. Our capture setup does not currently allow this, and image-based modelling of the tongue from photographs yields parameters poorly suited to animation. Were we to include head movements, eye blinks, and other nonarticulatory motions, this would inevitably lead to a great improvement in the naturalness of our output animations. Improvements could be achieved; yet the current system is focused upon creating natural lower facial for speech and would only be a part of a full facial animation system.

6. Evaluation

A short evaluation study has been conducted to determine the quality of the rendered animations. Seven subjects (with no special prior knowledge of the experimental setup) were shown synthetic sentences in several categories: (1) real data played back using the animation system (see Section 5); (2) animations generated using the model described in this paper; (3) animations generated using a technique which interpolates viseme centres. The interpolation method we use selects context-viseme examples from the dataset to match the phonetic transcription of the target utterance. These centres are interpolated using C^1 continuous Catmull-Rom splines to produce a continuous trajectory. The three different cases are each rendered using the same technique to remove any influence of the method of display on naturalness. Each animation consisted of three repetitions of a single sentence with natural audio, and the subject was asked to mark the quality of the animation on a 5-point scale from 1 (completely unnatural) through to 5 (completely natural). In total 66 sentences were presented to participants, 22 sentences repeated for each of the cases. The sentences selected for evaluation were taken from a 2-minute segment of recorded TIMIT sentences not used in training the model. These sentences were selected randomly and contained no overlap with the training set. The intention was to evaluate the quality of generated synthetic trajectories, whilst not also implicitly evaluating the quality of the animation technique itself. The playback of real data provides a ceiling on the attainable quality; that is, it is likely not possible to be more-real-than-real. Furthermore, the viseme-interpolation method is the lowest quality technique which does not produce entirely random or "babbling" speech animations. In this way we attempt to find where between these two quality bookends our technique falls. The results of the study for individual participants and overall are summarised in Table 3.

As expected overall and individually participants rated our method better than simple viseme interpolation. Generally, our technique came out as a mid-way point between the real and interpolated sentences. Furthermore, in some cases our technique was rated equal in quality to the equivalent animation from the real data, although this was for a minority of the sentences. The most obvious difference between our technique and the real motions is overarticulation. Our trajectories tend to articulate all the syllables in a sentence, whereas real speech tends to find a smoother trajectory. Having said this, our method does not overarticulate to the degree seen in the viseme-interpolation case, and the state-based model ensures that there is a strong constraint on how the lips move. Several subjects commented that the smoothness of the animation was a major factor in determining the naturalness of an animation. Potentially moving to a syllabic unit basis (or a multiscale basis, e.g., phoneme/triphone/syllable combined) may yield this smoothness, yet with the drawback of a much larger data capture requirement.

It is also worth noting that the results of our technique are quite variable, as is the case with most data-driven techniques. If an appropriate exemplar is not available in the database then the result can be a poor animation. It only takes a problem with a single syllable of a synthetic sentence to leave a large impact upon its perceived naturalness. Again this is most likely a problem of database size, notably audio speech synthesis databases are often far larger than the

8 minutes/103 sentences that we use as the basis for our system; however, the problem of capturing and processing a large corpus of visual speech movements needs to be solved to address this issue.

7. Summary and Discussion

In this paper we describe a hybrid technique for the synthesis of visual speech lip movements from audio, using elements of both unit selection and a global state-based model of speech movements. The underlying data for our system is captured surface movements for the lips and jaw gathered using a dynamic face capture system. By using dense surface data we are able to model the highly complex deformations of the lips during speech to a greater degree of accuracy than traditional capture techniques such as motion-capture and image-based modelling. From this data a speech manifold is recovered using dimensionality reduction techniques; this manifold demonstrates a strong structure related to the cyclical nature of speech lip movements. Our state-based model is constructed according to the clustering of data on this manifold. At synthesis time phonetic units are selected from the stored corpus and used to cull possible paths on the speech manifold and reduce the ambiguity in the mapping of audio speech parameters to visual speech lip movements. A *Viterbi*-type algorithm is used to determine an optimal traversal of the state-based model and infer a trajectory across the manifold and therefore a continuous sequence of lip movements. We generate animations using a layered model which combines the synthetic lip movements with a 3D jaw rotation model. The animations deform an image-plane according to the 3D speech lip movements and therefore create photorealistic output animations. A short perceptual study has been conducted to determine the quality of our output animations in comparison with both real data and simple viseme-interpolation. The results of this study indicate that in some cases our technique can be mistaken for real data (i.e., the naturalness is ranked equal or higher then the equivalent real movements), but in general the quality lies somewhere in-between the two extremes. In terms of evaluation this is not specific enough to truly define the quality of the technique, and further experimentation is required to compare with other existing techniques available in the literature.

The resulting animations are certainly far from perfect; we can see clearly from Figure 4 where the generated trajectory diverges from the real signal. It is worth noting that techniques driven entirely or partially (as is the case here) from audio tend to lag behind the quality of target driven techniques. This may be due to several factors, ranging from issues related to the capture of large visual speech databases to problems with the ambiguity in mapping from audio to visual trajectories. Visual speech databases, particularly in 3D, are far more difficult to capture than audio corpora. This is in large part due to the camera equipment used to capture facial movement, which in our case leads to restricted head movement (i.e., due to the size of the capture volume) and the need to place markers on the skin to get temporal registration. Any capture of this form is not going to get truly natural speech due to the intrusive nature of the setup, which may be a factor in the quality of our synthetic lip movements. Furthermore, the physical size of 3D databases and the time required to capture and reconstruct consistent data is a limiting factor in the size of our captured corpus. Eight minutes of data are small when compared to databases that are commonly used in speech analysis, and there is certainly an issue with sparsity when synthesising an utterance with our technique. With a data-driven approach missing data is a difficult problem to tackle, except with the obvious method of capturing more data. It is our hope that with the development of 3D capture technology these issues will be reduced, which will increase the viability of using surface capture technology for speech analysis and synthesis. Lastly, ambiguity in the mapping from audio to visual movements is also significant. We have found that it is generally true that clustering in the common audio parametric spaces (e.g., MFCC, PLP, etc.) does not lead to tight clusters in the visual domain, and vice versa when clustering in the visual domain. This is a fundamental problem and the motivation behind combining unit selection into the technique presented in this paper. However, this may be an issue with how we parameterise speech audio itself. These parametric spaces seem to serve speech recognition well, where we are decomposing a signal into a discrete sequence of symbols but may be less appropriate for generating continuous speech movements. There is a great deal of information within the audio signal which is not relevant to animating visual speech movements, for example, the distinction of nasalised or voiced sounds. There may also be information missing, such as information regarding respiration, which is important in producing realistic speech animations. It is obvious that the representation of the audio signal is key in determining the quality of animation from techniques such as our own, and perhaps research is required into the joint representation of speech audio and visual movements to reduce the ambiguity of this mapping.

Generating truly realistic speech animation is a very challenging task. The techniques described in this paper demonstrate the quality of animation that are attained when real lip movements can be used to infer the task space of speech production. Potentially capture techniques will advance such that more complex interactions between the lips and teeth can be captured (e.g., the f-tuck) which are not well modelled in the reported approach. However, this is only a part of the problem. To get truly natural characters we need to extend our models to full facial movement, to blinks, nods, and smiles. It is difficult to drive the movement of the articulators using the information embedded in a speech audio signal, let alone the complex emotional behaviour of a character. Yet this is the outcome that a viewer is looking for. Naturalness is perceived globally with regards to the movement of the entire face, and indeed body; this hampers current models which treat speech animation as an isolated part of human behaviour. It is probably the case that the next breakthrough in generating truly naturalistic synthetic facial animation will come as a result of a holistic approach to the modelling of behaviour, as opposed to

the piecemeal approaches commonly seen. Advances have currently been made as a result of data-driven modelling, as in this paper, and these approaches can yield convincing results. The drawback to such approaches lies in data capture; is it possible to capture truly comprehensive databases across speech and emotion? This is a huge problem that must be addressed if we are to reach the next level in purely synthetic character animation.

References

[1] M. Mori, "The uncanny valley," *Energy*, vol. 7, no. 4, pp. 33–35, 1970, translated by K. F. MacDorman and T. Minato.

[2] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp. 796–804, 1968.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '02)*, vol. 21, pp. 388–398, July 2002.

[4] I. Albrecht, J. Haber, and H.-P. Seidel, "Speech synchronization for physics-based facial animation," in *Proceedings of the 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG '02)*, pp. 9–16, 2002.

[5] L. Reveret, G. Bailly, and P. Badin, "Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. 755–758, 2000.

[6] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, Springer, Berlin, Germany, 1993.

[7] A. Löfqvist, "Speech as audible Gestures," in *Speech Production and Speech Modelling*, pp. 289–322, Springer, Berlin, Germany, 1990.

[8] M. Cohen, D. Massaro, and R. Clark, "Training a talking head," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, pp. 499–510, 2002.

[9] S. Öhman, "Numerical model of coarticulation," *Journal of the Acoustical Society of America*, vol. 41, pp. 310–320, 1967.

[10] Z. Deng, U. Neumann, J. P. Lewis, T.-Y. Kim, M. Bulut, and S. Narayanan, "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1523–1534, 2006.

[11] A. Black, P. Taylor, and R. Caley, "The festival speech synthesis system," 1999.

[12] T. Dutoit, V. Pagel, N. Pierret, E. Bataille, and O. van der Vrecken, "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, vol. 3, pp. 1393–1396, 1996.

[13] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: driving visual speech with audio," in *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics (SIGGRAPH '97)*, pp. 353–360, Los Angeles, Calif, USA, August 1997.

[14] Z. Deng and U. Neumann, "eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '06)*, pp. 251–260, 2006.

[15] S. Kshirsagar and N. Magnenat-Thalmann, "Visyllable based speech animation," in *Proceedings of the Annual Conference of the European Association for Computer Graphics (EUROGRAPHICS '03)*, vol. 22, pp. 631–639, September 2003.

[16] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, 2005.

[17] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[18] M. Brand, "Voice puppetry," in *Proceedings of the 26th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*, pp. 21–28, 1999.

[19] 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," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP '99)*, pp. 133–138, 1999.

[20] B. Theobald and N. Wilkinson, "A probabilistic trajectory synthesis system for synthesising visual speech," in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '08)*, 2008.

[21] T. Ezzat and T. Poggio, "Videorealistic talking faces: a morphing approach," in *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing (AVSP '97)*, pp. 141–144, 1997.

[22] J. D. Edge, A. Hilton, and P. Jackson, "Parameterisation of 3D speech lip movements," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP '08)*, 2008.

[23] P. Mueller, G. A. Kalberer, M. Proesmans, and L. Van Gool, "Realistic speech animation based on observed 3D face dynamics," *IEE Vision, Image & Signal Processing*, vol. 152, pp. 491–500, 2005.

[24] I. A. Ypsilos, A. Hilton, and S. Rowe, "Video-rate capture of dynamic face shape and appearance," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 117–122, May 2004.

[25] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz, "Spacetime faces: high resolution capture for modeling and animation," in *Proceedings of the 31st International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '04)*, pp. 548–558, Los Angeles, Calif, USA, August 2004.

[26] O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw, "A new trainable trajectory formation system for facial animation," in *Proceedings of the ISCA Workshop on Experimental Linguistics*, pp. 25–32, 2006.

[27] http://www.3dmd.com/.

[28] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *International Journal of Computer Vision*, vol. 13, no. 2, pp. 119–152, 1994.

[29] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proceedings of the DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.

[30] S. Roweis, "EM algorithms for PCA and SPCA," in *Proceedings of the Neural Information Processing Systems Conference (NIPS '97)*, pp. 626–632, 1997.

[31] J. Kruskal and M. Wish, *Multidimensional Scaling*, Sage, Beverly Hills, Calif, USA, 1979.

[32] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, pp. 374–388, Academic Press, New York, NY, USA, 1976.

[33] E. Vatikiotis-Bateson and D. J. Ostry, "Analysis and modeling of 3D jaw motion in speech and mastication," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 442–447, Tokyo, Japan, October 1999.

[34] D. J. Ostry, E. Vatikiotis-Bateson, and P. L. Gribble, "An examination of the degrees of freedom of human jaw motion in speech and mastication," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 6, pp. 1341–1351, 1997.

[35] E. Cosatto and H.-P. Graf, "Sample-based synthesis of photorealistic talking heads," in *Proceedings of the Computer Animation Conference*, pp. 103–110, 1998.

[36] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision (ECCV '98)*, pp. 484–498, 1998.