Animating Speech from Motion Fragments

James D. Edge Manuel A. Sánchez Lorenzo Steve Maddock

Technical Report CS-04-02
Department of Computer Science
University of Sheffield
UK

Abstract

The animation of facial expression has become a popular area of research in the past ten years, in particular with its application to avatar technology and naturalistic user interfaces. In this paper we describe a method to animate speech from small fragments of motion-captured sentences. A dataset of domain-specific sentences are captured and phonetically labelled, and from these sentences fragments are retrieved and blended to produce novel utterances. The movement of the motion-captured points is mapped onto a surface representation using a deformation technique based upon triangular Bézier patches. The resulting speech animation is highly realistic, and natural, preserving the correct articulatory effects expected from speech movements.

Key words: Facial Animation, Speech Synthesis, Free-form Deformation Techniques, Motion Capture

1 Introduction

Audio and visual stimuli are both involved in speech communication as a part of natural discourse. Not only does this involve emotional information, such as smiling or scowling, but also the movement of the lips, which is an important cue with regards the disambiguation of meaning. What we see and hear during speech gives complimentary information, which is backed up by perceptual studies that report as much as a +15dB improvement in signal-to-noise ratio [27] and a corresponding increase in the intelligibility of speech with visual information. This, along with interest in talking heads as a part of a more natural human-computer interface, has provoked a great deal of interest in synthesizing visual speech.

In this paper we discuss the synthesis of visual speech by concatenating short fragments from a library of motion-captured data. This idea is the analogue of the concatenative techniques used commonly in audio speech synthesis, allowing us to conceptually unify the models which deal with the audio and the visual streams. It is our assertion that speech animation in this manner is more natural and realistic than current popular techniques based upon the interpolation of visual phonemes.

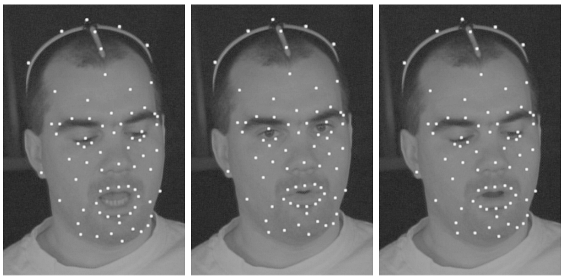


Figure 1: Frames from motion-captured sentences.

2 Background and Previous Work

In the following sections are reviews of the two main areas tackled in this paper: the modelling of facial expressions, and the animation of expression in synchrony with speech. For a detailed review of the entire area of facial animation, see [19].

2.1 Modelling Facial Expression

Techniques for the modelling of three-dimensional facial expressions fall into two broad categories: physical models of the skin and facial substructure, and free-form/geometric deformation. The first group of techniques embed the deformation within the face itself by attempting to accurately model the complex physical interaction between skin, muscles and the underlying boney substructure [21, 14, 12, 10]. Geometric techniques parameterize the deformation of a facial model based upon the location of a number of controlling structures, for example points, splines or patches placed upon its surface [18, 31, 11, 28, 23]. In both cases the intention is to provide realistic, fine-grained, and intuitive control over the surface of the skin to allow the synthesis of observed changes in facial expression.

Techniques for free-form deformation rely upon volumes defined by a few control primitives. These controlling primitives may be points [29], edges, splines [25], planes [26, 23], or volumes [24, 7, 16]. Bézier volume deformers have been used to animate faces both according to high level facial parameters [11], and to deform a facial mesh to match expressions in video sequences [28]. Radial basis functions have been used to interpolate the motion of a few points across a facial model [29]. In [23] triangular control elements, called Planar bones, are used to deform meshes for facial animation. The similarity between all these techniques is that a volume is being used to animate a surface, whereas a surface-to-surface deformation, like the BIDs deformation introduced in this paper, may be a more appropriate technique for facial animation.

2.2 Animating Speech

The animation of visible speech movements has lagged behind the corresponding techniques in audio speech synthesis. Much research in the topic of speech animation relies upon the simple interpolation between elementary speech units, often referred to as visual-phonemes or visemes. The problem in animating speech lies in the synchronicity of the visual movements with the audio and the naturalness of those movements. The naturalness of speech movements are judged against the experiences of the audience in real life, making the task much more difficult to solve than many areas in the field of animation.

One of the major difficulties in animating speech is the physical phenomenon called coarticulation [17, 15]. This term refers to the obscuration of boundaries between neighbouring atomic visual units. Whilst we may correctly be able to identify the lip shapes for each of the distinguishable sounds in a word, it is quite possible that none of these ideal targets will be met in natural discourse. Some of the targets are more important than others, and will be met to a greater or lesser extent accordingly. Furthermore, the degree to which each unit is met is coloured by its context, for example the articulation of the final /t/ in ’boot’ versus ’beet’. With this knowledge the most naïve methods to animate speech by direct interpolation of visemes are incorrect, and can result in visually disturbing movements.

Visible speech animation as a field focusses upon the recreation of coarticulation phenomena, and to this end several methods have been attempted: direct coarticulation modelling; mapping audio to visual parameters; and concatenation of visual units. The direct coarticulation models attempt to impose coarticulation upon the interpolation of atomic visual units. The second group of methods attempt to determine a direct relationship between audio and visual signals and exploit this in the synthesis of speech. Finally, concatenative methods take small chunks of real speech movements and paste them together to create novel utterances.

The most commonly used method for animating visual speech is to use dominance functions to represent the temporal extent of each atomic speech unit. This method, first proposed by Cohen and Massaro [6], has become the *de facto* standard for modelling coarticulation [9, 1]. Unfortunately, such systems require a high degree of tuning to create visually correct speech movement. Hidden Markov Models (HMMs) have been proposed [3] as a method for mapping between audio and visual parameters and thus drive the animation directly from the audio with no intermediate annotation of the speech. Also, highly complex models of the skin/muscle structure have been used in an inverse-dynamics approach to speech animation [20].

This paper describes a method for generating novel utterances using combinations of smaller speech fragments. The method is directly analogous to the most commonly used, and natural, audio synthesis methods which work by blending speech waveforms. Because the basis units come from real speech coarticulation is implicitly catered for within each unit. The challenges lie in correctly selecting and blending the units together to produce the appropriate visual movements in synchrony with the audio. Examples of concatenative visual synthesis include Video-Rewrite [4] which blends triphone video sequences, and more recent work by Kshirsagar [13] on using visual-syllables for synthesis.

3 Our Approach

In this paper we introduce a method for animating speech by concatenating small segments of natural speech movements. The original speech data is in the form of motion captured sentences, segmented into units of varying sizes according to phonetic structure (e.g. sentences, words etc.). By using a combination of natural motion fragments good quality speech animation can be achieved, without the necessity for complicated models of speech coarticulation.

Small fragments of speech are used to animate visual speech movements using a combination of motion warping, resampling and blending. Initially the fragments are warped such that they are phonetically aligned with the target utterance. Having stretched/squashed the fragments, each must be resampled to allow a consistent frame-rate throughout the animation. Finally, overlapping regions in the speech fragments are blended to provide smooth transitions.

In order to animate a high resolution model of the skin from the motion of a few sparsely sampled points an intermediate deformer surface is used to map the motion to individual vertices. This controlling structure is composed of a set of Bézier triangles spanning the motion captured points. The deformation technique provides smooth natural animation of a face mesh, even when only provided with a sparse sampling of the original facial motion.

The process can be summarised into the following stages:

- **Data Capture** - A corpus of natural human speech motion (visual component) and sounds (audio component) is captured.
- **Preprocessing** - Rigid motion and noise is removed from the motion-captured data.
- **Unit Generation** - Motion data is split into fragments representing the visual aspects of sentences, words and diphones (phone-to-phone transitions).
- **Motion Synthesis** - Combinations of motion fragments are used to generate visual information for novel speech utterances.
- **Animation** - Synchronized animation is produced using the BIDs (Bézier triangle Induced Deformations) technique to animate the visual component. The audio component is generated using the Festival speech synthesis system [2].

4 Data Description

The data used in this paper consists of motion data from a commercial Vicon capture system. High speed cameras, operating at 120Hz, capture the movement of 66 markers on the surface of an actors face plus 7 more on a head mounted jig to capture rigid motion (see fig. 1). Audio data was captured simultaneously and has been synchronized with the motion data (see fig. 2).

Fifty-five sentences were captured from a limited domain time corpus, the sentences take the following form:

prompt	:=	{*prolog*} / {*time-info*} / {*day-info*}.
time-info	:=	{*exactness*} {*minutes*} {*hours*}
prolog	:=	'the time is now'
exactness	:=	'exactly' **or** 'just after' **or** 'a little after' **or** 'almost'
minutes	:=	'five past' **or** 'ten past' **or** 'quarter past' **or** 'twenty past' **or** 'twenty-five past' **or** 'half past' **or** 'twenty-five to' **or** 'twenty to' **or** 'quarter to' **or** 'ten to' **or** 'five to'
hours	:=	'one' **or** 'two' **or** . . . **or** 'twelve'
day-info	:=	'in the morning' **or** 'afternoon' **or** 'am' **or** 'pm'

This corpus can be used to generate simple time sentences such as:

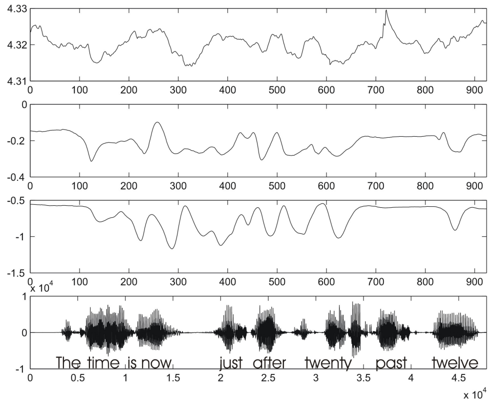


Figure 2: Animation parameters and speech waveform for a sample captured sentence.

'the time is now / exactly one / in the afternoon.'
or *'the time is now / quarter to ten / in the morning.'*

The data is specific to the time domain, and thus the implemented system presented in this paper is limited in generality. However, the techniques described are equally applicable to larger corpora or general synthesis using, for example, diphones as the lowest level speech unit. A simple corpus has been used to demonstrate the general technique and to ensure consistency in the dataset.

The captured motions require some processing in order to both remove noise and reconstruct missing data. Kalman filtering is used to remove noise from the data, whilst resampling of the DCT is used to reconstruct the missing data segments, typically caused by marker occlusion. The rigid head motion is also removed at this stage using a combination of the estimate from the head mounted jig and a least-squares approach. This last step has the added benefit that the motion samples are initially spatially aligned enabling simpler concatenation during synthesis.

5 Animation from Facial Motion Capture

The final synthesis of visible speech movements in this paper relies upon the ability to animate a high resolution facial model directly from the motion of a few motion-captured control points. Two stages are required to animate in this manner: firstly the motion data must be retargetted such that it is embedded within the target facial model; secondly, the motion of these points must be mapped onto the higher resolution surface to provide a continuous and natural deformation. The solutions to these problems are detailed in the next sections.

5.1 Retargetting Motions

The retargetting problem for facial motion capture is analogous to similar problems with full-body mocap [32]. Given differences in facial shape and scale captured motions must be modified to make them applicable to the animation of any individual mesh. Far less work has been conducted into the case of facial motion capture than for full-body. Here the technique from [23] is used to retarget motions such that they become embedded within a given target mesh.

The technique relies upon defining a mapping between a frame in the source motion and the target surface. Given such a mapping, points in subsequent frames can be retargetted by maintaining their relative position. A volume warp is used to provide this mapping. Once this warp is established for one frame of the motion, subsequent frames can be retargetted by its reapplication. The use of such a warp is correct for all motions with the constraint that they can only exhibit small movement perpendicular to the surface, which we find is the case for facial motion.

The retargetting requires only the labelling of a few points on the target geometry, but is otherwise completely automatic. It is beyond the scope of this paper to describe the entire technique in detail, however, it is important to translate motions into the space of the target mesh before animation. Please refer to the original paper [23] for details of the retargetting.

5.2 Bézier-triangle Induced Deformations (BIDs)

Facial motion capture data consists only of the motion of a few sparse points on the surface of the skin. In order to animate a target mesh from the motion of these points it is necessary to interpolate the displacements across the facial surface. The BIDs deformation technique constructs a surface as a triangulation of the control points, and proceeds by defining a one-to-one mapping from vertices in the target surface onto the deformer surface. Due to the fact that the deformer itself is composed of Bézier triangles, a certain degree of surface and deformation continuity can be maintained. Two stages are required by the BIDs deformation technique: projection, and reconstruction.

Projection

In order to define a parameterization of the target geometry each vertex is projected onto the deformer surface along the surface normal of the closest point (see fig. 3). This is possible because the deformer surface is constrained such that its unitary normal map is at least C_0 continuous, and provides a full coverage of the target mesh.

For a vertex, V, the closest point in the parametric domain of the i^{th} Bézier triangle can be found by minimizing the square distance. This problem is equivalent to finding the root of (1), where B_i is the biparametric expression of the i^{th} triangular patch.

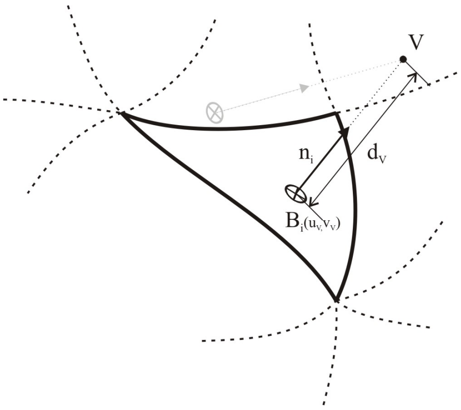


Figure 3: Projection of vertex V onto the closest triangular patch B_i. Another candidate over an adjacent Bézier triangle is shown in light gray.

$$\left(\frac{\partial\|V - B_i(u,v)\|^2}{\partial u}\right)^2 + \left(\frac{\partial\|V - B_i(u,v)\|^2}{\partial v}\right)^2 = 0 \qquad (1)$$

Solutions may be found for $\{u_V, v_V\}$ that fall out of the parametric domain of the patch, however, taking into account the coverage property previously introduced, it can be stated that at least one of the tuples found will lie inside its respective domain triangle.

It must be noted that in the case of cubic and quartic Bézier surfaces, the degree of this expansion prevents us from using an analytical solution. In these cases a numerical approach must be taken. Due to the regularity of low-degree Bézier triangles a Newton-Raphson gradient descent approach should be capable of finding the correct minima of (1). To ensure convergence the starting point for the optimization procedure is derived from a coarse sampling of the surface.

To complete the parameterization for a vertex, V, the projected distance along the surface normal, d_V, is stored. Given a parameterization, $\{u_V, v_V, d_V\}$, each of the target vertices can be directly reconstructed from the deformer surface as described in the next section.

Reconstruction

The second step of the algorithm makes use of the defined parameterization to reconstruct the deformed target geometry over the control surface spanned by the displaced control points (e.g. displaced in subsequent frames of animation). This is the inverse process to the projection described in the previous section. The deformation of a vertex, V, projected onto the i^{th} Bézier triangle is defined in (2).

$$\begin{aligned} def(V) &= B'_i(u_V, v_V) + d_V n'_i(u_V, v_V) \\ \text{where} & \\ n'_i(u,v) &= \hat{N}'_i(u,v) \\ N'_i(u,v) &= \frac{\partial B'_i}{\partial u}(\mathbf{u},\mathbf{v}) \wedge \frac{\partial B'_i}{\partial v}(\mathbf{u},\mathbf{v}) \end{aligned} \qquad (2)$$

In (2) B'_i and n'_i are the deformed Bézier triangle and normal respectively, i.e. deformed according to the displaced control points and additional continuity considerations. Thus, given a parameterization of a target surface, deformed surfaces can simply be reconstructed by re-evaluating the necessary Bézier triangle patches.

Enforcing Continuity

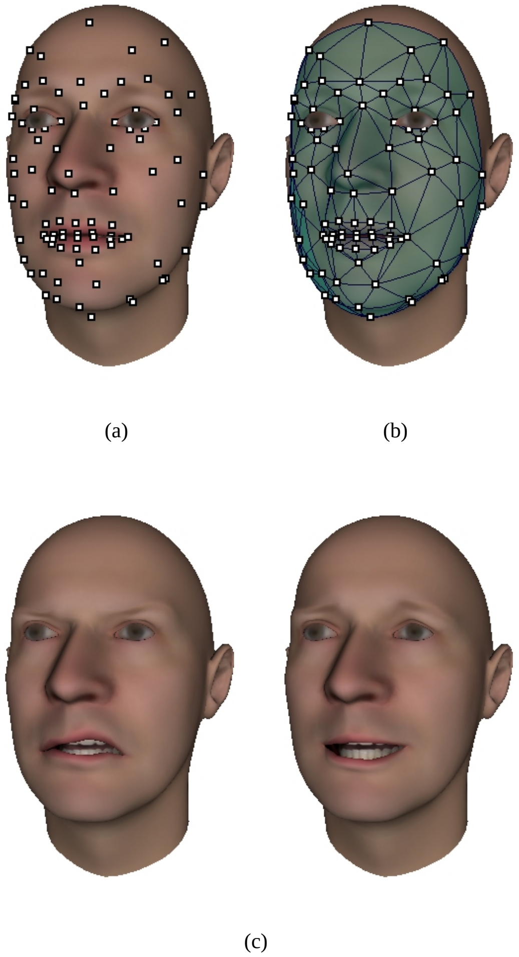
Preserving the smoothness of the target surface is key to producing natural animation, especially in the case of animating faces. In the case of the BIDs technique the continuity of the deformation is directly related to that enforced upon the control surface. If the control surface has G^n geometric continuity, then the deformation will be C^{n-1} continuous. This is the case except for non-bijective singularities, in these rare occurances the choice of attachment is made according to the orientation of the surface (see Surface Discontinuities).

In order to ensure at least C^0 continuity, the deformer surface must be constrained so that the continuity across patch boundaries is G^1. This is achieved by performing Clough-Tocher partitioning into cubic sub-patches [5], and then conditioning the resulting control net in a similar manner to Veltkamp [30]. This conditioning scheme can be performed for every triangle independently of the control points in its neighborhood, which enables *locality* in the deformation. This is because the shape of a given Bèzier patch is only affected by the deformation of control points on the adjacent triangle-patches. Higher degree continuity in the deformation, e.g. C^1, would require greater geometric continuity in the deformer surface and appropriate conditioning methods. For further discussion of geometric continuity see [8].

Surface Discontinuities

In order to preserve skin discontinuities, for example between the lips, a strategy must be adopted to control the attachment of vertices to the deformer surface. This is particularly important in the case that the boundaries of the deformer surface do not directly match the target surface, which will always be the case as the control points are sparsely sampled.

Masks are often used to determine which deformers affect which areas of the target surface [23]. However, in the case of the BIDs technique, because there is a direct surface-to-surface correspondance, a comparison of the deformer surface and vertex normals can be used to correctly determine attachment. The assumption is that the deformer surface should be a reasonable approximation of the target, and thus where attachment is ambiguous surface orientation is a good disambiguating feature. Thus, when two patches are similarly distant from a vertex the attachment is chosen which is closest in orientation to the vertex.


(a)

(b)

(c)

Figure 4: Face control structure: (a) motion-captured points; (b) triangulated Bézier control surface; (c) modelled facial expressions.

6 Speech Animation

The vast majority of techniques for the synthesis of visual speech movement rely upon the interpolation of a set of atomic phonetic units. The more successful paradigm, certainly in the case of audio synthesis, relies upon the concatenation of natural segments of speech. The analogous methods in the visual domain are becoming more popular [4, 13]. In this paper fragments of visual speech are concatenated to provide speech animation.

6.1 Visual Speech Fragments

As previously mentioned animating speech from small motion fragments provides the advantage that coarticulation need not be modelled, and the naturalness of the movements is implicit. However, there are also problems with this data-driven approach. Primarily, a database covering the entirity of the target domain must be captured. This impinges upon the size of fragments captured, for example if diphone (phone-to-phone) transitions are used there will be approximately 1500 units for British English. Larger units, such as syllables and words, will require an even greater (possibly unmanagable) database for synthesis. This choice of synthesis unit is a matter of balance, as it is also the case that larger fragments produce more natural animation.

Here, for the purposes of demonstration, sentences from the time domain are used. From these sentences diphone, syllable, word, and sentence fragments are extracted for synthesis. Together these fragments can be used to resynthesize any sentence from the time domain described in Section 4. In order to construct novel utterances from these fragments the following stages must be conducted:

- **Unit Selection** - Appropriate units must be selected from the database to generate the utterance.
- **Phonetic Alignment** - Each of the selected units must be phonetically aligned such that the movements appear in synchrony with the speech.
- **Resampling** - As a consequence of alignment speech fragments must be resampled to a consistent frame-rate for animation.
- **Blending** - Having aligned and resampled the motions, overlapping sections are blended to achieve a consistent trajectory over the synthesized utterance.
- **Retargetting and Animation** - A target face model is animated from the synthesized speech movements using the techniques in Section 5.

Unit Selection

The technique for unit selection used is dependent upon the underlying speech units. In this case units of varying duration are available, and thus a method must be defined to select the most appropriate selection to synthesize a target utterance. As input to the process the phonetic labels and timing of the target utterance are required, which can be directly recovered from the audio synthesis procedure (in this case the Festival synthesis system [2]). Pseudocode for the basic algorithm is shown below.

```
Fragment Selection Algorithm
Input: List of $phones$
Output: List of $fragments$

$frags \leftarrow []$
$i \leftarrow 1$
$j \leftarrow numPhones$
while $i < numPhones$ do
    while not FIND-UNIT$(phones, i, j)$ do
        $j \leftarrow j - 1$
    end while
    APPEND-UNIT$(frags, phones, i, j)$
    $i \leftarrow j$
    $j \leftarrow numPhones$
end while
```

In this code FIND-UNIT is a subprocedure which searches for a speech fragment which spans several phones in the target utterance, e.g. the closed sequence ['c','a','t']. APPEND-UNIT appends the found unit to the output list of fragments. Primarily this algorithm chooses fragments of longer duration, which is beneficial to the naturalness of the output speech. However, disambiguation is required where more than one speech fragment is available within the database for a given sequence. In this case, the factors which are taken into account when selecting units are: similarity in the phonetic timing to the target utterance, and similarity of context. Each of these conditions biases towards using fragments as similar as possible to the target utterance, and thus the synthesized trajectories should maintain the naturalness in movement of the captured data.

Alignment and Resampling of Speech Fragments

Given an appropriate selection of units, the next stage is to adapt these fragments so that in combination they can be used to synthesize the target utterance. Essentially, this requires that the units are temporally aligned with the target utterance. Each speech fragment, whether it be diphone or a sentence, has a phonetic labelling, and must be variously stretched/squashed so that the labels are correctly aligned with the phonetic structure of the synthesized audio.

Simply, this can be achieved by evenly distributing motion samples between repositioned phonetic labels. However, this will lead to an uneven distribution in the sampling of the speech fragments, which will give an inconsistent frame-rate for animation. For this reason, having adapted the fragments so that they are aligned with the target utterance, the fragments must be further resampled to achieve a consistent frame-rate before blending.

This is the scattered-data interpolation problem, i.e. given a scattered sampling of data form a continuous curve/surface passing through the points. Many methods, such as B-spline interpolation, could be used to resample the data, here radial-basis functions (RBFs) are used.

The RBF method forms an interpolant as a linear combination of basis functions (3).

$$f(x) = p_m(x) + \sum_{i=1}^{n} \alpha_i \phi(|x - c_i|) \tag{3}$$

In (3) the interpolated point, $f(x)$, is a linear combination of n basis functions, $\phi(x)$, and a polynomial term, $p_m(x)$. Each basis function is termed *radial* because its scalar value depends only upon the distance from its centre, c_i. The basis function used here is the inverse multiquadric, which has the advantage of being continuous in all derivatives, i.e. C^{∞}. The key step in using this form of interpolation is to determine the weights, α_i, which ensure that all of the basis centres are exactly interpolated. The weights can simply be determined by placing the basis centres back into (3), and solving the resulting system of linear equations. For a more thorough discussion of RBF interpolation refer to [22].

To use RBFs for the purposes of resampling motion fragments, a basis centre is placed at each sampled point, ensuring that the interpolating curve will exactly fit the known data. The interpolated motions are in fact a mapping from the time-domain onto the spatial domain, and thus to finally resample the data requires only querying the interpolated motion at uniform temporal intervals.

Blending Motions

The final stage of synthesis, given appropriate aligned speech fragments from the previous stages, is to blend the fragments such that continuous motion is exhibited in the resulting animation. This involves only the overlapping regions of motions at the joints, a small degree of context is required in the fragments to facilitate this. Within the overlapping section, $t \in [t_0, t_1]$, a weighted blend of the two motion fragments to be concatenated is used (4).

$$\theta_{blend}(t) = g(u)\theta_0(t) + (1 - g(u))\theta_1(t) \tag{4}$$

$$where \quad u = \left(\frac{t - t_0}{t_1 - t_0}\right)$$

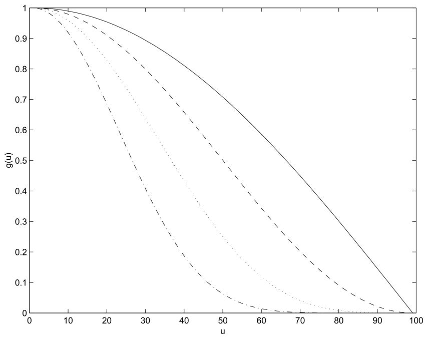


Figure 5: Example weighting functions, $g(u)$.

In (4), $g(u)$ is a weighting function (see fig. 5) which returns a value in the interval $[0, 1]$. The weighting function facilitates the blend and ensures a smooth transition between the fragments, which are represented here as functions of time ($\theta_x(t)$). The speed of decay in g will determine how fast the second fragment is faded in.

The use of blending relies upon the alignment of the motion fragments which is ensured in a preprocessing stage along with the removal of extraneous noise in the signals. The size of the overlapping regions depends upon the frame-rate of the fragments themselves, however, they should always be a fraction of the smallest phone-to-phone interval to prevent large fragments dominating over the target utterance. In practice, for animation frame-rates of 30 fps, there will never be more than a couple of frames overlap at each join, and for this reason high speed capture is advantageous as it allows larger blend intervals.

7 Results and Conclusions

The previous sections have described a system for the animation of speech by concatenating small fragments of motion captured data. Several frames from an animation produced using this technique are shown in Figure 6. The system currently implemented is limited to the domain of time sentences, however, the techniques described are equally applicable to larger scale corpora and general synthesis where appropriate datasets are available. The advantages of concatenative synthesis, over interpolative techniques in particular, lies in the naturalness of the movements which can be achieved. To produce equivalent animations using conventional synthesis techniques would require a great deal of effort, particularly in modelling coarticulation. Furthermore, the described system takes advantage of a retargetting technique [23] to allow motions gathered from a single actor to be reused multiple times to animate facial meshes varying in both shape and scale.

We have introduced a novel deformation technique (BIDs) based upon a mapping between a Bézier triangle control surface and the geometry of a mesh. BIDs produces continuous deformation according to the motion of a few control points. The technique itself is generic and could be used for any free-form deformation task, however, we find that it is particularly appropriate to the case of facial animation as the deformer surface can interpolate the motion-captured points and map the movement directly onto an underlying facial mesh. The BIDs deformer surface used in the example animations is shown in Figure 4.

It is evident from recent interest in the speech synthesis community [4, 13] that concatenative synthesis shall soon become a predominant technique for the animation of visual speech. As the equipment to capture facial movement becomes cheaper and more accessible it is likely that motion databases will be increasingly used for animation. Data-driven approaches have thus far been far more successful than pure synthesis for audio. It is likely that the same will ultimately be true in the visual domain.

Acknowledgements

The authors would like to thank the EPSRC and the Pedro Barrié de la Maza Foundation for providing funding for this research, as well as the Advanced Computing Center for the Arts and Design Motion Capture Lab for the collection of the motion data. We would also like to thank Dr Scott King for his help in producing this research.

References

[1] I. Albrecht, J. Haber, and H-P Seidel. Speech synchronization for physics-based facial animation. In *Proceedings WSCG'02*, pages 9–16, 2002.

[2] A. Black, P. Taylor, and R. Caley. The festival speech synthesis system. (`http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html`), June 1999.

[3] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999.

[4] C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997.

[5] R. Clough and J. Tocher. Finite element stiffness matrices for analysis of plates in bending. In *Proceedings of Conference on Matrix Methods in Structural Analysis*, 1965.

[6] M.M. Cohen and D.W. Massaro. Modeling coarticulation in synthetic visual speech. In *Proceedings Computer Animation '93*, pages 139–156, 1993.

[7] S. Coquillart. Extended free-form deformation: a sculpturing tool for 3d geometric modeling. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, pages 187–196, 1990.

[8] G. Farin. *Curves and Surfaces for Computer Aided Geometric Design. Fourth ed.* Academic Press, 1997.

[9] B. Le Goff and C. Benoît. A text-to-audiovisual-speech synthesizer for french. In *Proc. ICSLP'96*, volume 4, pages 2163–2166, Philadelphia, PA, 1996.

[10] K. Kähler, J. Haber, and H-P. Siedel. Geometry-based muscle modeling for facial animation. In *Proceedings Graphics Interface 2001*, pages 37–46, 2001.

[11] P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. Simulation of facial muscle actions based on rational free form deformations. In *Proceedings Eurographics'92*, pages 59–69, 1992.

[12] R.M. Koch, M.H. Gross, and A.A. Bosshard. Emotion editing using finite elements. In *Proceedings Eurographics'98*, 1998.

[13] S. Kshirsagar and N. Magnenat-Thalmann. Visyllable based speech animation. In *Proceedings Eurographics 2003*, 2003.

[14] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62, 1995.

[15] A. Löfqvist. Speech as audible gestures. *Speech Production and Speech Modelling*, pages 289–322, 1990.

[16] R. MacCracken and K.I. Joy. Free-form deformations with lattices of arbitrary topology. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 181–188, 1996.

[17] S.E.G. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310–320, 1967.

[18] F.I. Parke. *A parametric model for human faces*. PhD thesis, University of Utah, 1974.

[19] F.I. Parke and K. Waters. *Computer Facial Animation*. A. K. Peters, Ltd., 1996.

[20] M. Pitermann and K.G. Munhall. An inverse dynamics approach to face animation. *Journal of the Acoustical Society of America*, 110:1570–1580, 2001.

[21] S.M. Platt and N.I. Badler. Animating facial expressions. In *Proceedings of the 8th annual conference on Computer graphics and interactive techniques*, pages 245–252, 1981.

[22] D. Ruprecht and H. Muller. Image warping with scattered data interpolation. *IEEE Computer Graphics and Applications*, 3:37–43, 1995.

[23] M. Sánchez, J.D. Edge, S.A. King, and S. Maddock. Use and re-use of facial motion capture data. In *Proceedings Vision, Video and Graphics*, pages 135–142, 2003.

[24] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 151–160, 1986.

[25] K. Singh and E. Fiume. Wires: a geometric deformation technique. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 405–414, 1998.

[26] K. Singh and E. Kokkevis. Skinning characters using surface oriented free-form deformations. In *Proceedings Graphics Interface*, pages 35–42, 2000.

[27] W.H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

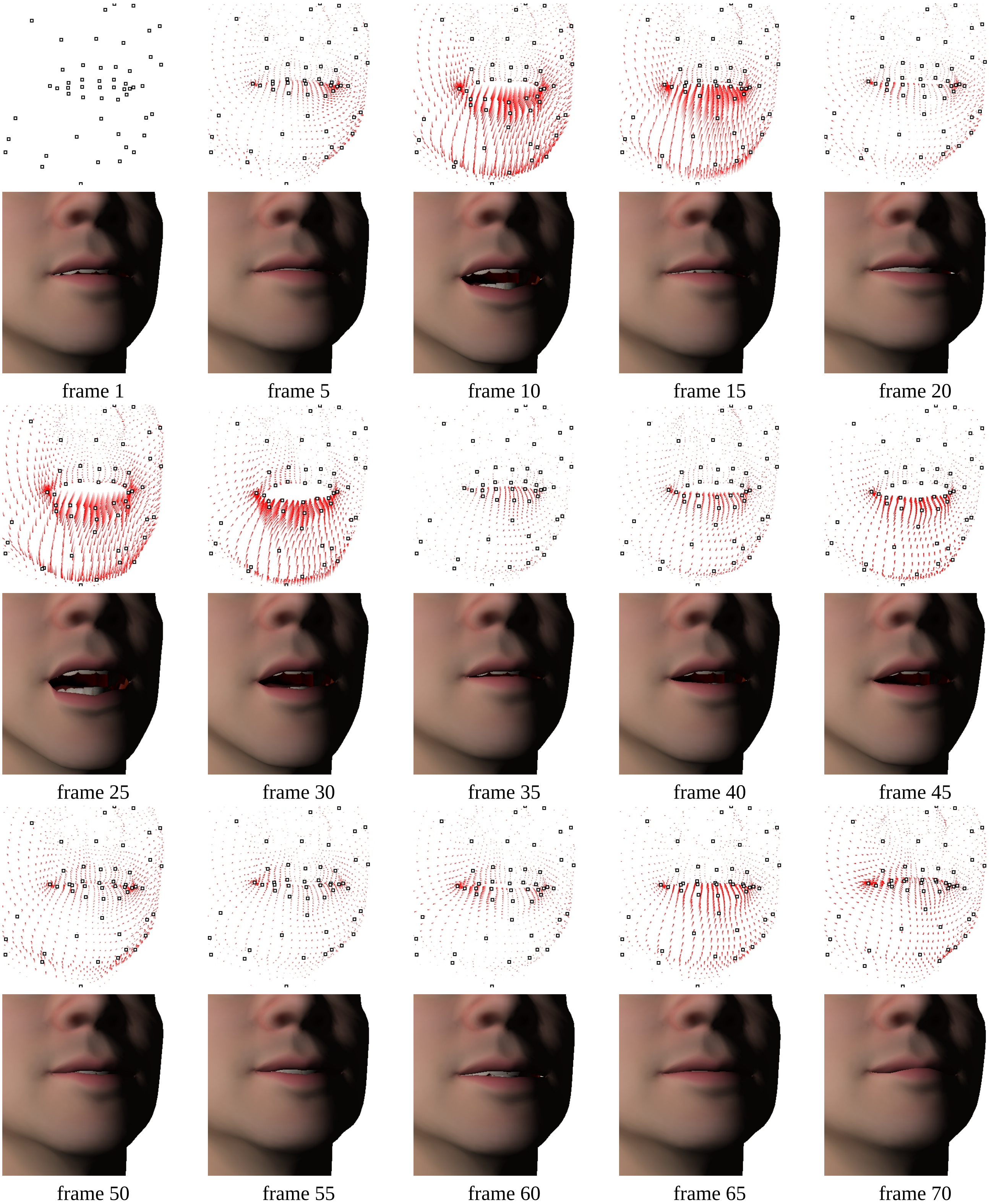
[28] H. Tao and T. Huang. Bezier volume deformation model for facial animation and video tracking. In *Proceedings CAPTECH'98*, pages 242–253, 1998.

[29] F. Ulgen. A step toward universal facial animation via volume morphing. In *Proceedings 6th IEEE International Workshop on Robot and Human communication*, pages 358–363, 1997.

[30] R.C. Veltkamp. Closed G^1-continuous cubic bézier surfaces. Technical report, University of Utrecht, 1992.

[31] K. Waters. A muscle model for animation three-dimensional facial expression. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 17–24, 1987.

[32] A. Witkin and Z. Popovic. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 105–108, 1995.



frame 1 frame 5 frame 10 frame 15 frame 20

frame 25 frame 30 frame 35 frame 40 frame 45

frame 50 frame 55 frame 60 frame 65 frame 70

Figure 6: Example frames and vertex trajectories from a speech animation.