Reusing Motion Data to Animate Visual Speech

James D. Edge Manuel A. Sánchez Lorenzo Steve Maddock

*University of Sheffield
Regent Court
211 Portobello st
Sheffield S1 4DP
j.edge@dcs.shef.ac.uk
m.sanchez@dcs.shef.ac.uk
s.maddock@dcs.shef.ac.uk

Abstract

In this paper we describe a technique for animating visual speech by concatenating small fragments of speech movements. The technique is analogous to the most effective audio synthesis techniques which form utterances by blending small fragments of speech waveforms. Motion and audio data is initially captured to cover the target synthesis domain; this data is subsequently phonetically labelled and segmented to provide basis units for synthesis. Sentence, word and syllable level units are used by the system to synthesize novel speech utterances. The final synthesized utterances consist of the motion of points on the surface of the skin, these trajectories are retargetted and interpolated across the surface of a target mesh for animation.

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 [Sumby and Pollack (1954)] 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.

Visual speech synthesis is a sub-field within general modelling and animation of facial expression. In this context expression is the grouping of gestural (e.g. conversational signals), emotional (e.g. scowling) and physical (e.g. blinking) actions required to communicate between individuals. Speech is inherently a physical process of shaping the vocal tract such that a meaningful sequence of sounds is created, according to the words and grammar of a particular language. This is particularly important because the physical relationship between the visual and audio modalities necessitates that facial movements we observe are 'correct'; where this is not the case perceptual difficulties may arise [McGurk and MacDonald (1976)] or at least there will be an impediment to the realism of the animation.

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.

2 Background

Much research into facial animation considers the difficulties in modelling static facial expression [Parke (1974); Waters (1987); Lee et al. (1995)]. Here we review the more specialized field of speech animation, for a detailed overview of the entire field see [Parke and Waters (1996)]. For a detailed discussion of visual speech synthesis and relating perceptual issues see [Massaro (1998)].

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 [Öhman (1967); Löfqvist (1990)]. 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 [Cohen and Massaro (1993)], has become the *de facto* standard for modelling coarticulation [Goff and Benoît (1996); King et al. (2000); Revéret et al. (2000); Albrecht et al. (2002)]. Unfortunately, such systems require a high degree of tuning to create visually correct speech movement. For example in [Cohen and Massaro (1993)] they require three parameters per function/parameter pair plus a global parameter controlling the shape of the dominance functions. Hidden Markov Models (HMMs) have been proposed [Brooke and Scott (1998),Brand (1999)] as a method for mapping between audio and visual parameters and thus to 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 [Pitermann and Munhall (2001)].

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 [Bregler et al. (1997)] which blends triphone video sequences, and more recent work by Kshirsagar [Kshirsagar and Magnenat-Thalmann (2003)] on using visual-syllables (visyllables) for synthesis.

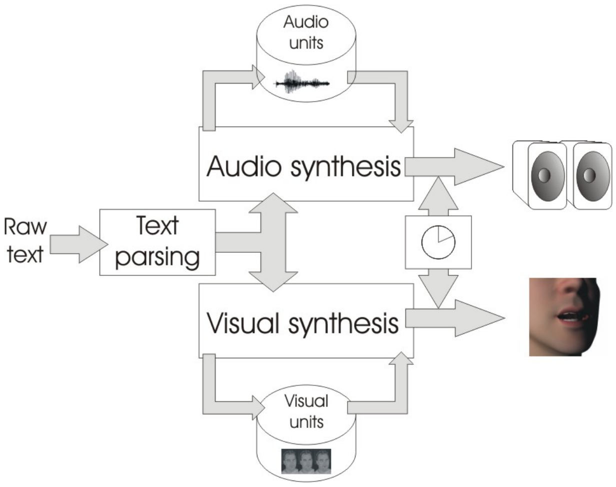


Figure 1: Overview of the synthesis system.

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. phones, syllables, 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.

Motion data is captured using a commercial Vicon mocap system. High-speed cameras, operating at 120Hz, capture the movement of markers placed on an actors face. The resulting data is a sparse sampling of the surface motion of the skin during speech production.

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.

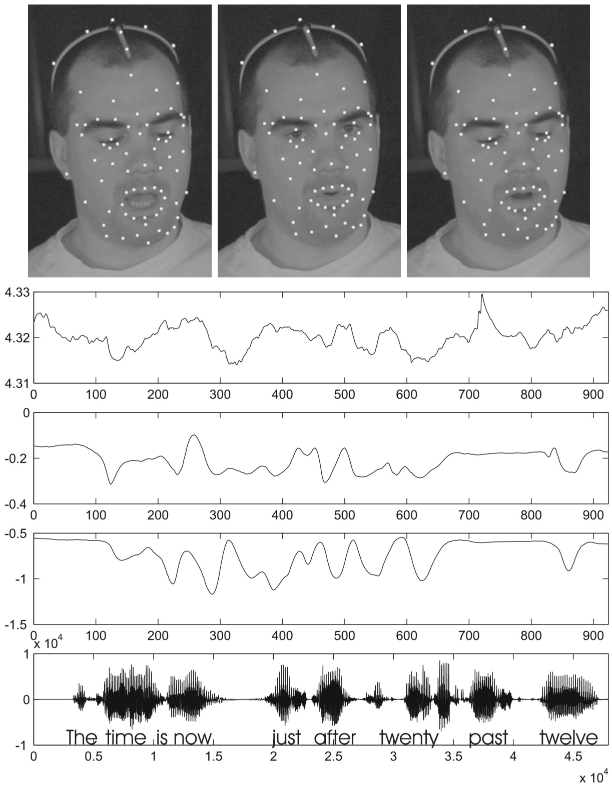


Figure 2: Example frames, motion trajectories and waveform from the captured database.

- **Preprocessing** - Rigid motion (e.g. head 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.
- **Retargetting** - Synthesized motions are transformed into the space of the target mesh according to the algorithms described in [Sánchez et al. (2003)]. This allows the captured motion data to be used in the animation of meshes which vary in both shape and scale from the original actor.
- **Animation** - Synchronized animation is produced using the BIDs (Bézier triangle Induced Deformations) [Edge et al. (2004)] technique to animate the visual component.

The implemented system performs Text-to-Audio-Visual-Speech (TTAVS) synthesis; an overview of the system structure is shown in Figure 1. Synthesis is split directly into separate audio and visual components: Festival [Black et al. (1999)] is used to synthesize the audio; visual synthesis is the focus of this paper and consists of unit selection, temporal alignment, and blending to generate the final motion. Both audio and visual synthesis components hold databases of speech data, synchronously captured from the same actor so that the synthesized motions will be correct for the synthesized audio. As input to the system raw text is taken from the user, this is phonetically annotated to act as input to the process. The same input is used for both audio and visual synthesis, however, each system is fundamentally independant and there is no assertion that the corresponding units will be used in both modalities.

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. Audio data was captured simultaneously and has been synchronized with the motion data. Figure 2 shows several frames from the captured data alongside captured motion parameters and audio.

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:

'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 Speech Synthesis

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 [Bregler et al. (1997); Kshirsagar and Magnenat-Thalmann (2003)]. In this paper fragments of visual speech are concatenated to provide speech animation.

5.1 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 6.

5.1.1 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 [Black et al. (1999)]). 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.

5.1.2 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 (1).

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

In (1) 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 (1), and solving the resulting system of linear equations. For a more thorough discussion of RBF interpolation refer to [Ruprecht and Muller (1995)].

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.

5.1.3 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 (2).

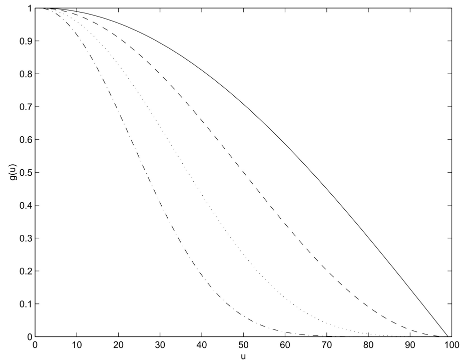


Figure 3: Example weighting functions.

$$\theta_{blend}(t) = g(u)\theta_0(t) + (1 - g(u))\theta_1(t) \qquad (2)$$

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

In (2), $g(u)$ is a weighting function (see fig. 3) 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.

6 Retargetting and Animation

The result of synthesis by the techniques described in this paper leads to motion trajectories for a sparse sampling of points on the source actors face. This data is limited without further processing both to retarget the motions to a particular target mesh and to embed the motion in that mesh by interpolating the motion of points across its surface.

In order to retarget the motion to a target mesh we use the method described in [Sánchez et al. (2003)]. Because a mesh may vary in shape, scale and orientation this method consists of a volume warping method which provides a continuous mapping from the space of the original motion captured data to the space of the target mesh, i.e. $f : \mathbb{R}^3 \rightarrow \mathbb{R}^3$.

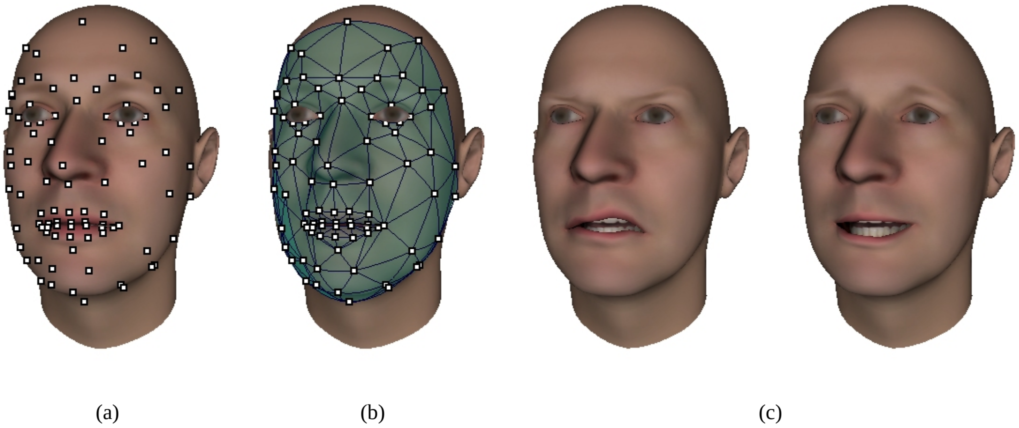


Figure 4: Face control structure: (a) motion-captured points; (b) triangulated Bézier control surface; (c) modelled facial expressions.

The mapping is determined using RBFs, described earlier in Section 5.1.2, to define a mapping between a neutral facial pose in the source motion and the target mesh. By reapplication of the warping function subsequent frames in the motion can be retargetted relative to the neutral pose. The fundamental technique is relatively simple, however, its application requires several technical issues to be addressed. For a full discussion of the retargetting task for facial animation and details of the technique in particular refer to the original paper [Sánchez et al. (2003)].

The result of the retargetting process is a sparsely-sampled motion embedded in a target mesh. Given that the mesh itself is often far more densely sampled than the motion it is important to interpolate the motion across the surface in a manner that preserves both the motion itself, and the characteristic geometric structure of the mesh (for example, the discontinuity between the lips).

In the system a surface oriented free-form deformation technique is used for this purpose. Free-form deformation tools provide control over high resolution meshes using a small number of controlling structures, usually lattices of control points [Sederberg and Parry (1986); Coquillart (1990); Singh and Fiume (1998)]. Here a deformer surface is defined as a triangulation of the motion captured points (fig. 4 (a) and (b)). This controlling structure is a Bézier triangle surface with continuity conditions at patch boundaries. Vertices in the target mesh are parameterized according to the parametric coordinates, $[u_V, v_V]$, of their projected image on the closest controller element (Bézier triangle), along with a normal offset, d_V, from the surface (3).

$$V_{def} = B_i(u_V, v_V) + d_V n_i(u_V, v_V) \qquad (3)$$

In (3), B_i is the parametric definition of the i^{th} triangular Bézier patch, and n_i its unitary normal map. As control points in the deformer surface are manipulated the target mesh will deform accordingly, maintaining its geometric relationship with the deformer. Figure 4 (c) shows example modelled facial expressions created using this technique.

Further constraints can be placed upon the attachment process to maintain discontinuities in the target mesh. This consists of thresholding the maximum angle allowed between the surface normals of the vertex and its image in the parametric domain of the closest Bézier element. Such constraints assert a similarity condition for the attachment of the target to the deformer. This is particularly important in controlling the movement of the lips which must be able to move entirely independently. Also, this technique does not require any form of explicit masking [Sánchez et al. (2003)] or other manual labour to be applied to an entirely different mesh.

The deformation technique is used to interpolate the movement of motion-captured points across the surface of a target mesh. Because the deformer surface approximates the mesh we achieve realistic and physically plausible movement from only a sparse sampling of an actors face. A more detailed description of the Bézier Induced Deformations (BIDs) technique can be found in [Edge et al. (2004)].

7 Results

Several frames and motion trajectories from an example animation are shown in Figure 5. Animations generated by the system demonstrate physically plausible motions, as would be expected given that the basis-units for synthesis are directly captured movements. This is particularly evident in the movement of the skin in the cheeks which is not often accounted for in morphable models of vocal articulation. The skin visibly stretches and bulges as you would expect from a physical model of the human skin, e.g. [Lee et al. (1995)].

The described system is only capable of deriving skin movement, which is the only movement evident in the initial motion captured samples. In order to animate the tongue and lower jaw either a more complex database needs to be captured (e.g. using electropalatography to determine tongue movement) or, as has been done here, a backoff technique can be used. The movement of the jaw is determined directly from the motion of captured points on the skin covering the jaw using. The tongue uses a simple morph-based model which is adequate given that it is often occluded both by the teeth and the lips.

All animation techniques described in this paper can be implemented in real-time on current PC hardware. The synthesis of individual phrases is also not particularly computationally intensive, however, the processing time will necesarily depend upon the length of the target utterance and the number of motion fragments required. Preprocessing and data preparation tasks can be labour intensive (for example, phonetically labelling the captured audio), but only need to be performed once per database.

8 Conclusions

There are three key advantages to using motion fragments in visual speech synthesis:

- Motion-captured data implicitly encapsulates dynamic coarticulation effects.
- It allows the unification of audio and visual synthesis by using the correct motions for the audio units concatenated during synthesis.
- An improvement in the naturalness of speech movements is attained, especially in comparison to interpolation techniques such as [Cohen and Massaro (1993); Goff and Benoît (1996); Albrecht et al. (2002)].

Furthermore, due to the use of retargetting and generic animation techniques, the implemented system is capable of driving any reasonable facial mesh. The parameters used are not model-specific, nor are we constrained to use particular point-sets (e.g. MPEG-4), or mesh topologies (e.g. [Parke (1974)]) making the system both generic and scalable.

One improvement over the currently implemented system would be to link unit selection in the audio and visual modalities. For each motion sequence we also retain the audio data, which is used by Festival for synthesis. Currently there is no link between unit selection, and so visual units may be selected which were not captured at the same time as the selected audio units. Using Festival to select both audio *and* visual units may remedy this and produce a slight benefit in audio-visual synchronicity. Furthermore, there may be some benefit in expanding upon selection criteria for visual units to bias towards units with given boundary conditions, i.e. similarity in the overlapping blend period.

The disadvantages to using motion capture for these purposes lies in the size of databases required to perform general synthesis (as opposed to limited-domain, e.g. rail announcements or time - as in this paper). Larger motion units will lead to an increase in the quality of synthesized speech, however, it also leads to larger-scale initial data capture. For example, using triphones as the lowest level speech unit will require approximately 1500 units in British English. Capturing this amount of audio-visual data is highly complex, particularly with regards to maintaining consistency in the database. In order to tackle these problems we use multiple scale units (phones, syllables, words etc.) to allow the greatest quality synthesis for the captured data. By the aforementioned use of retargetting techniques we are also able to capture once and use the data multiple times, i.e. to animate several characters.

One direction for future research is the use of backoff techniques to merge concatenative techniques with morph based models. This is analogous to the letter-to-sound rules used in audio synthesis and would allow motion-data to be used even without restricting synthesis to domain-specific problems or requiring large-scale data capture. Currently no systems have attempted to mix synthesis techniques in this manner, and the most commonly used coarticulation technique [Cohen and Massaro (1993)] is inappropriate for the task because continuity considerations at utterance boundaries are not taken into account.

The future of visual speech synthesis lies in the use of larger dynamic units, as has been long recognised by the audio synthesis community. The remaining problems focus upon concurent signals such as emotional and gestural signals. This would require that algorithms are developed to blend sampled motions without destroying the link between speech movements and audio. The authors believe that the wealth of research in full-body motion capture [Bruderlin and Williams (1995)] as well as recent developments in decomposing motions [Cao et al. (2003)] could be exploited to tackle this problem.

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

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

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.

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

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.

N.M. Brooke and S.D. Scott. Two and three-dimensional audio visual speech synthesis. *Proc. AVSP'98*, 1998.

A. Bruderlin and L. Williams. Motion signal processing. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 97–104, 1995.

Y. Cao, P. Faloutsos, and F. Pighin. Unsupervised learning for speech motion editing. In *Proceedings of Symposium on Computer Animation*, pages 225–231, 2003.

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

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.

J.D. Edge, M.A. Sánchez, and S. Maddock. Animating speech from motion fragments. Technical Report CS-04-02, Department of Computer Science, University of Sheffield, 2004.

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.

S. King, R.E. Parent, and B.L. Olafsky. An anotomically-based 3d parametric lip model to support facial animation and synchronized speech. *Proc. Deform 2000*, pages 7–9, 2000.

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

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.

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

D. Massaro. *Perceiving Talking Faces : From Speech Perception to a Behavioral Principle.* Bradford Books Series in Cognitive Psychology. MIT Press, 1998.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

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

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

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

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

Lionel Revéret, Gérard Bailly, and Pierre Badin. Mother : A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. *6th Int. Conference of Spoken Language Processing, ICSLP'2000*, 2000.

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

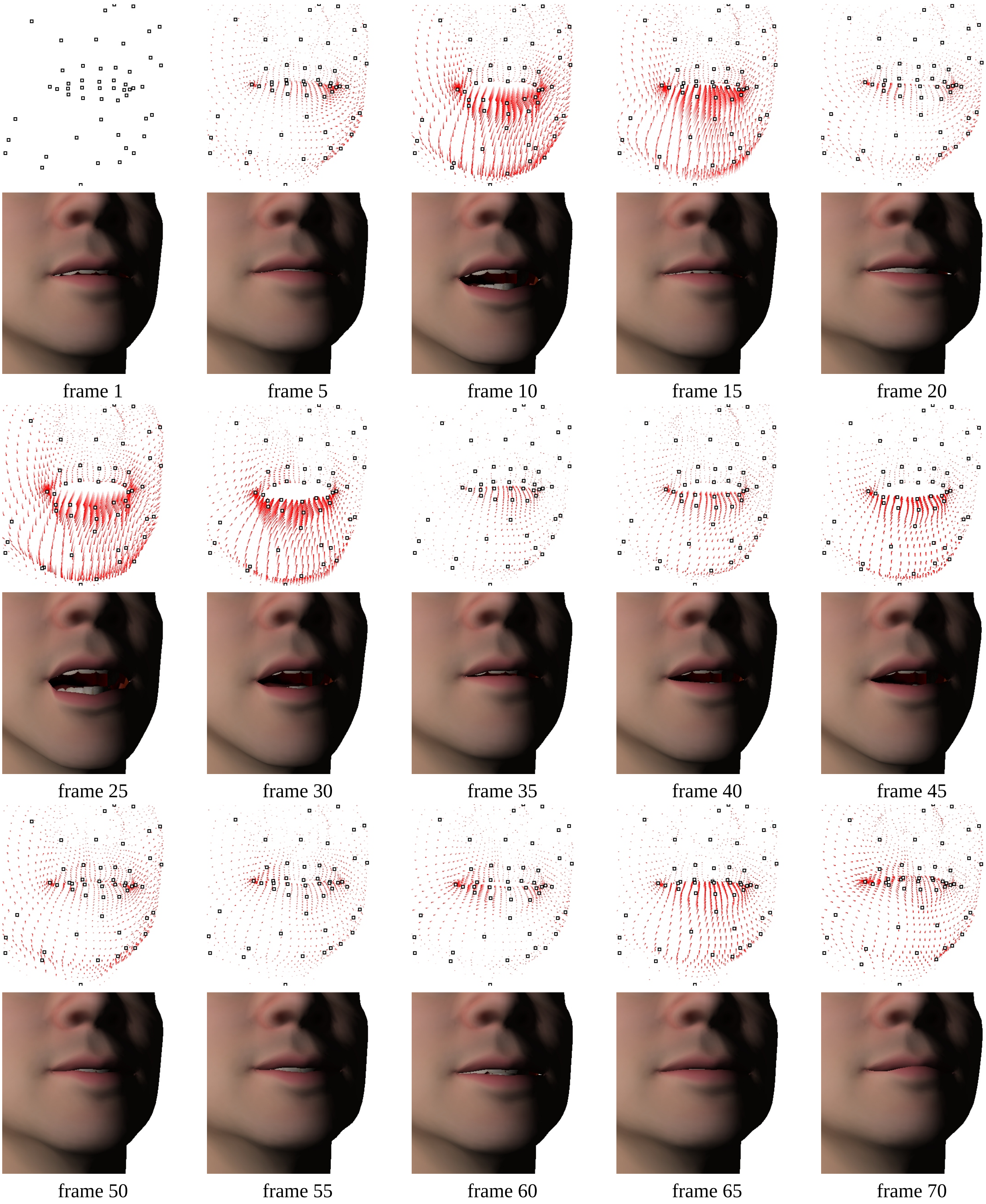
M. Sánchez, J.D. Edge, S.A. King, and S. Maddock. Use and re-use of facial motion capture data. In *Proceedings Vision, Video and Graphics*, pages 135–142, 2003.

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.

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.

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.

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.



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 5: Example frames and vertex trajectories from a speech animation.