Hindawi Publishing Corporation
International Journal of Computer Games Technology
Volume 2008, Article ID 412056, 7 pages
doi:10.1155/2008/412056

Research Article

A Constraint-Based Approach to Visual Speech for a Mexican-Spanish Talking Head

Oscar Martinez Lazalde, Steve Maddock, and Michael Meredith

Department of Computer Science, Faculty of Engineering, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK

Correspondence should be addressed to Oscar Martinez Lazalde, acp03om@sheffield.ac.uk

Received 30 September 2007; Accepted 21 December 2007

Recommended by Kok Wai Wong

A common approach to produce visual speech is to interpolate the parameters describing a sequence of mouth shapes, known as visemes, where a viseme corresponds to a phoneme in an utterance. The interpolation process must consider the issue of context-dependent shape, or coarticulation, in order to produce realistic-looking speech. We describe an approach to such pose-based interpolation that deals with coarticulation using a constraint-based technique. This is demonstrated using a Mexican-Spanish talking head, which can vary its speed of talking and produce coarticulation effects.

1. INTRODUCTION

Film, computer games, and anthropometric interfaces need facial animation, of which a key component is visual speech. Approaches to producing this animation include pose-based interpolation, concatenation of dynamic units, and physically based modeling (see [1] for a review). Our approach is based on pose-based interpolation, where the parameters describing a sequence of facial postures are interpolated to produce animation. For general facial animation, this approach gives artists close control over the final result; and for visual speech, it fits easily with the phoneme-based approach to producing speech. However, it is important that the interpolation process produces the effects observed in the natural visual speech. Instead of treating the pose-based approach as a purely parametric interpolation, we base the interpolation on a system of constraints on the shape and movement of the visible parts of the articulatory system (i.e., lips, teeth/jaw, and tongue).

In the typical approach to producing visual speech, the speech is first broken into a sequence of phonemes (with timing), then these are matched to their equivalent visemes (where a viseme is the shape and position of the articulatory system at its visual extent for a particular phoneme in the target language, e.g., the lips would be set in a pouted and rounded position for the /u/ in "b**oo**"), and then intermediate poses are produced using parametric interpolation. With less than sixty phonemes needed in English, which can be mapped onto fewer visemes since, for example, the bilabial plosives /p/, /b/, and the bilabial nasal /m/ are visually the same (as the tongue cannot be seen in these visemes), the general technique is low on data requirements. Of course, extra postures would be required for further facial postures such as expressions or eyebrow movements.

To produce good visual speech, the interpolation process must cater for the effect known as coarticulation [2], essentially context-dependent shape. As an example of forward coarticulation, the lips will round in anticipation of pronouncing the /u/ of the word "stew," thus affecting the articulatory gestures for "s" and "t." The de facto approach used in visual speech synthesis to model coarticulation is to use dominance curves [3]. However, this approach has a number of problems (see [4] for a detailed discussion), perhaps the most fundamental of which is that it does not address the issues that cause coarticulation.

Coarticulation is potentially due to both a mental planning activity and the physical constraints of the articulatory system. We may plan to over- or underarticulate, and we may try to, say, speak fast, with the result that the articulators cannot realize their ideal target positions. Our approach

tries to capture the essence of this. We use a constraint-based approach to visual speech (first proposed in [4, 5]), which is based on Witkin and Kass's work on physics-based articulated body motion [6]. In [7], we presented the basics of our approach. Here, we show how it can be used to produce controllable visual speech effects, whilst varying the speed of speech.

Section 2 will present an overview of the constraint-based approach. Sections 3, 4, and 5 demonstrate how the approach is used to create Mexican-Spanish visual speech for a synthetic 3D head. Section 3 outlines the required input data and observations for the constraint-based approach. Section 4 describes the complete system. Section 5 shows the results from a synthetic talking head. Finally, Section 6 presents conclusions.

2. CONSTRAINT-BASED VISUAL SPEECH

A posture (viseme) for a phoneme is variable within and between speakers. It is affected by context (the so-called coarticulation effect), as well as by such things as mood and tiredness. This variability needs to be encoded within the model. Thus, a viseme is regarded as a distribution around an ideal target. The aim is to hit the target, but the realization is that most average speakers do not achieve this. Highly deformable visemes, such as an open mouthed /a/, are regarded as having larger distributions than closed-lip shapes, such as /m/. Each distribution is regarded as a constraint which must be satisfied by any final speech trajectory. As long as the trajectory stays within the limits of each viseme, it is regarded as acceptable, and infinite variety within acceptable limits is possible.

To prevent the ideal targets from being met by the trajectory, other constraints must be present. For example, a global constraint can be used to limit the acceleration and deceleration of a trajectory. In practice, the global constraint and the distribution (or range) constraints produce an equilibrium, where they are both satisfied. Variations can be used to give different trajectories. For example, low values of the global constraint (together with relaxed range constraints) could be used to simulate underarticulation (e.g., mumbling). In addition, a weighting factor can be introduced to change the importance of a particular viseme relative to others.

Using the constraints and the weights, an optimization function is used to create a trajectory that tries to pass close to the center of each viseme. Figure 1 gives a conceptual view of this. We believe that this approach better matches the mental and physical activity that produces the coarticulation effect, thus leading to better visual speech. In using a constrained optimization approach [8], we need two parts: an objective function $\mathrm{Obj}(X)$ and a set of bounded constraints C_j,

$$\text{minimize } \mathrm{Obj}(X) \quad \text{subject to } \forall j : \underline{b}_j \le C_j(X) \le \overline{b}_j, \quad (1)$$

where $\underline{b}_j$ and $\overline{b}_j$ are the lower and upper bounds, respectively. The objective function specifies the goodness of the system state X for each step in an iterative optimization procedure. The constraints maintain the physicality of the motion.

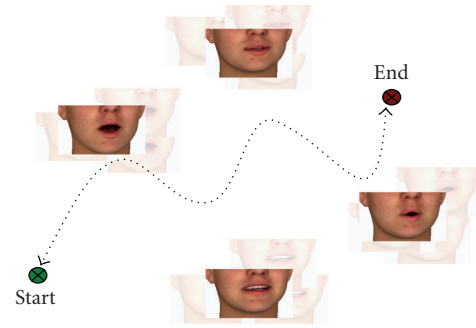


Figure 1: Conceptual view of the interpolation process through or near to clusters of acceptable mouth shapes for each viseme.

Table 1: Boundary constraints.

Constraints	Action
$S(t_{\text{start}}) = \varepsilon_{\text{start}}$	Ensures trajectory starts at $\varepsilon_{\text{start}}$
$S(t_{\text{end}}) = \varepsilon_{\text{end}}$	Ensures trajectory ends at ε_{end}
$S(t_{\text{start}})' = S(t_{\text{end}})' = 0$	Ensures the velocity is equal to zero at the beginning and end of the trajectory
$S(t_{\text{start}})'' = S(t_{\text{end}})'' = 0$	Ensures the acceleration is equal to zero at the beginning and end of the trajectory

The following mathematics is described in detail in [4]. Only a summary is offered here. The particular optimization function we use is

$$\mathrm{Obj}(X) = \sum_i w_i \left(S(t_i) - V_i\right)^2. \quad (2)$$

The objective function uses the square difference between the speech trajectory S and the sequence of ideal targets (visemes) V_i, given at times t_i. The weights w_i are used to give control over how much a target is favored. Essentially, this governs how much a target dominates its neighbors. Note that in the presence of no constraints, w_i will have no impact, and the V_i will be interpolated.

A speech trajectory S will start and end with particular constraints, for example, a neutral state such as silence. These are the boundary constraints, as listed in Table 1, which ensure the articulators in the rest state. If necessary, these constraints can also be used to join trajectories together.

In addition, range constraints can be used to ensure that the trajectory stays within a certain distance of each target,

$$S(t_i) \in [\underline{V}_i, \overline{V}_i], \quad (3)$$

where $\underline{V}$ and $\overline{V}_i$ are, respectively, the lower and upper bounds of the ideal targets V_i.

If (3) and Table 1 are used in (2), the ideal targets V_i will simply be met. A global constraint can be used to dampen the trajectory. We limit the parametric acceleration of a trajectory.

$$\left|S(t)''\right| \le \gamma, \quad \text{where } t \in [t_{\text{start}}, t_{\text{end}}], \quad (4)$$

Table 2: Mexican-Spanish viseme definitions.

Phoneme	Viseme name	Phoneme	Viseme name
Silence	Neutral	i	I
j, g	J	c, k, q	K
b, m, p, v	B_M_P	n, ñ	N
a	A	o,u	O
ch, ll, y, x	CH_Y	r	R
d, s, t, z	D_S_T	l	L

and γ is the maximum allowable magnitude of acceleration across the entire trajectory. As this value tends to zero, the trajectory cannot meet its targets, and thus the w_i in (2) begins to have an effect. The trajectory bends more towards the target, where w_i is high relative to its neighbors. As the global constraint is reduced, the trajectory will eventually reach the limit of at least one range constraint.

The speech trajectory S is represented by a cubic nonuniform B-spline. This gives the necessary C^2 continuity to enable (4) to be applied. The optimization problem is solved using a variant of the sequential quadratic programming (SQP) method (see [6]). The SQP algorithm requires the objective function described in (2). It also requires the derivatives of the objective and the constraints functions: the Hessian of the objective function H_{obj} and the Jacobian of the constraints J_{cstr}. This algorithm follows an iterative process with the steps described in (5). The iterative process finishes when the constraints are met, and there is no further reduction in the optimization function (see Section 5 for discussion of this):

$$\Delta X_{\text{obj}} = -H_{\text{obj}}^{-1} \begin{pmatrix} \dfrac{\partial \text{Obj}}{\partial X_1} \\ \vdots \\ \dfrac{\partial \text{Obj}}{\partial X_n} \end{pmatrix},$$
$$\Delta X_{\text{cstr}} = -J_{\text{cstr}}^{+}(J_{\text{cstr}} \Delta X_{\text{obj}} + C), \tag{5}$$
$$X_{j+1} = X_j + (\Delta X_{\text{obj}} + \Delta X_{\text{cstr}}).$$

3. INPUT DATA FOR THE RANGE CONSTRAINTS

In order to produce specific values for the range constraints described in Section 2, we need to define the visemes that are to be used and measure their visual shapes on real speakers. In English, there is no formal agreement on the number of visemes to use. For example, Massaro defines 17 visemes [9], and both Dodd and Campbell [10], as well as Tekalp and Ostermann [11] use 14 visemes. We chose 15 visemes for Mexican-Spanish, as listed in Table 2.

Many of the 15 visemes we chose are similar to the English visemes, although there are exceptions. The phoneme /v/ is an example, where there is a different mapping between Spanish and English visemes. In English speech, the phoneme maps to the /F/ viseme, whereas in Spanish, the /v/ phoneme corresponds to the /B_M_P/ viseme. There are also letters, like /h/, that do not have a corresponding phoneme in Spanish (they are not pronounced during speech) and thus have no associated viseme. Similarly, there are phonemes in Spanish that do not occur in English, such as /ñ/, although there is an appropriate viseme mapping in this example to the /N/ viseme.

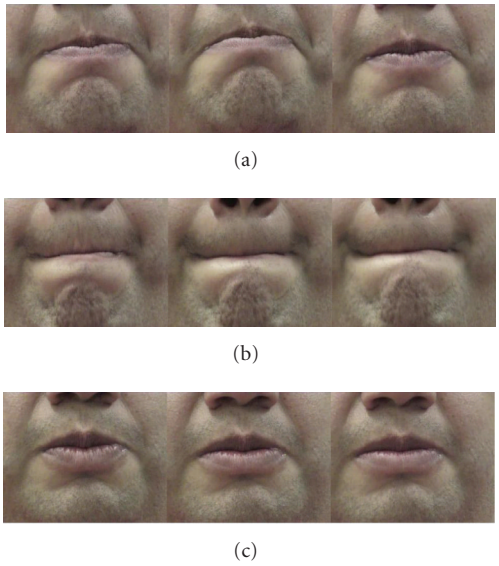
Figure 2: The left two columns show the front and side views of the viseme M. The right two columns show the front and side views of the viseme A. (a) The synthetic face; (b) Person A; (c) Person B; (d) Person C.

To create the range constraints for the Mexican-Spanish visemes listed in Table 2, three native Mexican-Spanish speakers were observed, labeled Person A, Person B, and Person C. Each was asked to make the ideal viseme shapes in Mexican-Spanish, and these were photographed from front and side views. Figure 2 gives examples of the lip shapes for the consonant M (labelled as B_M_P in Table 2) and for the vowel A for each speaker, as well as the modeled synthetic head (which was produced using FaceGen www.facegen.com). Figure 3 shows the variation in the lip shape for the consonant M when Person B pronounces the word "ama" normally, with emphasis and in a mumbling style. This variation is accommodated by defining upper and lower values for the range constraints. Figure 4 illustrates the issue of coarticulation. Person B was recorded three times pronouncing the words "ama," "eme," and "omo," and the frames containing the center of the phoneme "m" were extracted. Figure 4 shows that the shape of the mouth is more rounded in the pronunciation of "omo" because the phoneme m is surrounded by the rounded vowel o.

4. THE SYSTEM

Figure 5 illustrates the complete system for the Mexican-Spanish talking head. The main C++ module is in charge of communication between the rest of the modules. This module first receives text as input, and then gets the corresponding phonetic transcription, audio wave, and timing from a Festival server [12]. The phonetic transcription is used to retrieve the relevant viseme data. Using the information from Festival together with the viseme data, the optimization problem is defined and passed to a MATLAB routine, which contains the SQP implementation. This returns a spline definition and the main C++ module, then generates the rendering of the 3D face in synchronization with the audio wave.

Each viseme is represented by a 3D polygon mesh containing 1504 vertices. Instead of using the optimization process on each vertex, the amount of data is reduced using

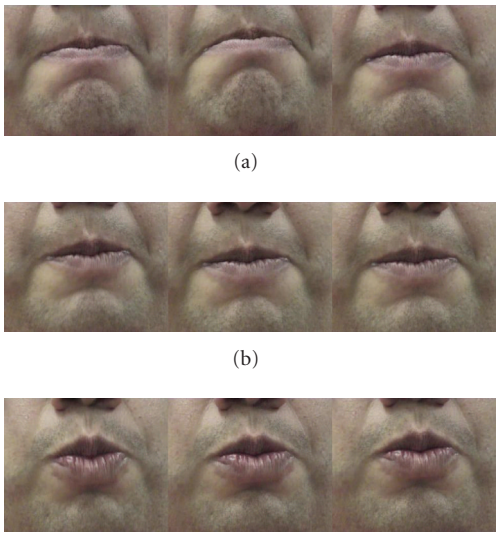


(a)

(b)

(c)

Figure 3: Visual differences in the pronunciation of the phoneme m in the word "ama": (a) normal pronunciation; (b) with emphasis; (c) mumbling. In each case, Person B pronounced the word 3 times to show potential variation.



(a)

(b)

(c)

Figure 4: Contextual differences in the pronunciation of the phonene m: (a) the m of "ama"; (b) the m of "eme"; (c) the m of "omo." In each case, Person B pronounced the word 3 times to show potential.

principal component analysis (PCA). This technique reconstructs a vector V_i that belongs to a randomly sampled vector population V using (6)

$$V = \{v_0, v_1, \ldots, v_s\},$$
$$v_i = u_V + \sum_{j=1}^{s} e_j b_j, \quad 0 \le j \le s, \tag{6}$$

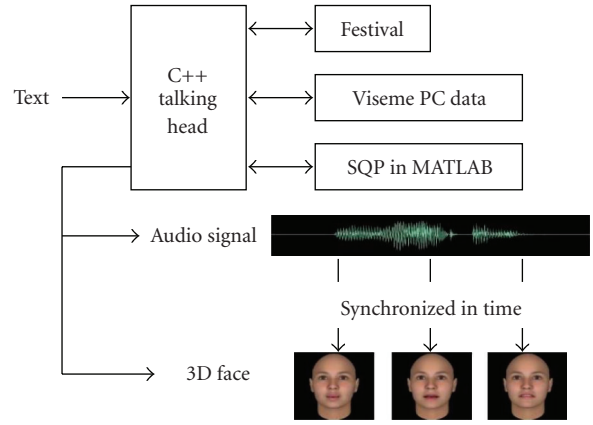


Figure 5: Talking-head system.

where u_V is the mean vector, e_i are the eigenvectors obtained after applying the PCA technique, and b_j are the weight values. With this technique, it is possible to reconstruct, at the cost of minimal error, any of the vectors in the population using a reduced number of eigenvectors e_j and its corresponding weights b_j.

To do the reconstruction, all the vectors share the reduced set of eigenvectors e_j (PCs), but they use different weights b_j for each of those eigenvectors. Thus, each viseme is represented by a vector of weight values.

With this technique, the potential optimization calculations for 1504 vertices are reduced to calculations for a much smaller number of weights. We chose 8 PCs by observing the differences between the original mesh and the reconstructed mesh using different numbers of PCs. Other researchers have used principal components as a parameterization too, although the number used varies from model to model. For example, Edge uses 10 principal components [4], and Kshirsagar et al. have used 7 [13], 8 [14], and 9 [15] components.

It is the PCs that are the parameters (targets) that need to be interpolated in our approach. In the results section, we focus on the PC 1, which relates to the degree that the mouth is open. To determine the range constraints for this PC, the captured visemes were ordered according to the amount of mouth opening. Using this viseme order, the range constraint values were set accordingly using a relative scale. The same range constraint values were set for all other PCs for all visemes. Whilst PC 2 does influence the amount of mouth rounding, we decided to focus on PC 1 to illustrate our approach. Other PCs only give subtle mouth shape differences and are difficult to determine manually. We hope to address this by working on measuring range constraints for static visemes using continuous speaker video. The acceleration constraint is also set for each PC.

5. RESULTS

The Mexican-Spanish talking head was tested with the sentence "hola, cómo estas?". Figure 6 shows the results of the

mouth shape at the time of pronouncing each phoneme in the sentence. Figures 7 and 8 illustrate what is happening for the first PC in producing the results of Figure 6. The pink curves in Figures 7 and 8 show that the global constraint value is set high enough so that all the ideal targets (mouth shapes) are met (visual results in Figure 6(a)). Figure 6(b) and the blue curves in Figures 7 and 8 illustrate what happens when the global constraint is reduced. In Figure 8, the acceleration (blue curve) is restricted by the global acceleration constraint (horizontal blue line). Thus, the blue spline curve in Figure 7 does not meet the ideal targets. Thus, some of the mouth shapes in Figure 6(b) are restricted. The more notable differences are at the second row (phoneme l), at the fifth row (phoneme o), and at the tenth row (phoneme t).

In each of the previous examples, both the global constraint and the range constraint could be satisfied. Making the global constraint smaller could, however, lead to an unstable system, where the two kinds of constraints are "fighting." In an unstable system, it is impossible to find a solution that satisfies both kinds of constraints; and as a result, the system jumps from a solution that satisfies the global constraint to one that satisfies the range constraint in an undetermined way leading to no convergence. To make the system stable under such conditions, there are two options: relax the range constraints or relax the global constraint. The decision on what constraint to relax will depend on what kind of animation is wanted. If we were interested in preserving speaker-dependent animation, we would relax the global constraints as the range constraints encode the boundaries of the manner of articulation of that speaker. If we were interested in producing mumbling effects or producing animation where we were not interested in preserving the speaker's manner of articulation, then the range constraint could be relaxed.

Figure 6(c) and the green curves in Figures 7 and 8 illustrate what happens when the global constraint was reduced further so as to make the system unstable, and the range constraints were relaxed to produce stability again. In Figure 7, the green curve does not satisfy the original range constraints (solid red lines), but does satisfy the relaxed range constraints (dotted red lines). Visual differences can be observed in Figure 6 at the second row (phoneme l), where the mouth is less open in Figure 6(c) than in Figures 6(a) and 6(b). This is also apparent at the fifth row (phoneme o) and at the tenth row (phoneme t).

For Figure 6(d), the speed of speaking was decreased resulting in a doubling of the time taken to say the test sentence. The global constraint was set at the same value as for Figure 6(c), but this time the range constraints were not relaxed. However, the change in speaking speed means that the constraints have time to be satisfied as illustrated in Figures 9 and 10 .

As a final comment, the shape of any facial pose in the animation sequence will be most influenced by its closest visemes. The nature of the constraint-based approach means that the neighborhood of influence includes all visemes, but is at its strongest within a region of 1-2 visemes, either side of the facial pose being considered. This range corresponds to most common coarticulation effects, although contextual effects have been observed up to 7 visemes away [16].

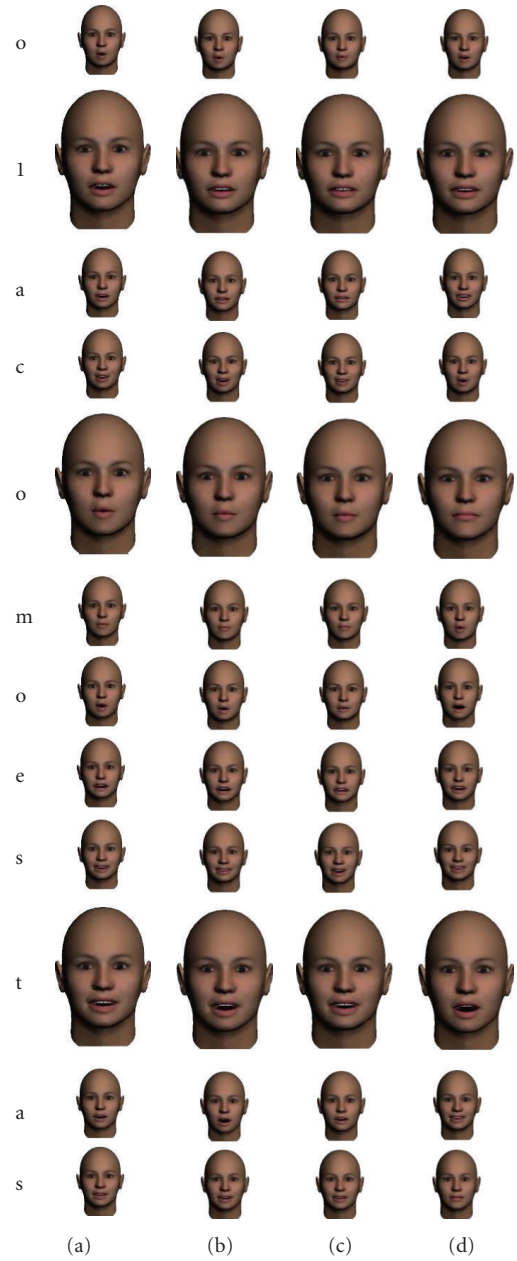


Figure 6: Face positions for the sentence "hola, cómo estas?": (a) targets are met (global constraint 0.03); (b) targets not met (global constraint 0.004); (c) targets not met (global constraint 0.002) and range constraints relaxed; (d) speaking slowly and targets not met (global constraint 0.002).

6. CONCLUSIONS

We have produced a Mexican-Spanish talking head that uses a constraint-based approach to create realistic-looking speech trajectories. The approach accommodates speaker variability and the pronunciation variability of an individual speaker, and produces coarticulation effects. We have demonstrated this variability by altering the global constraint, relaxing the range constraints, and changing the speed of speaking. Currently, PCA is employed to reduce the

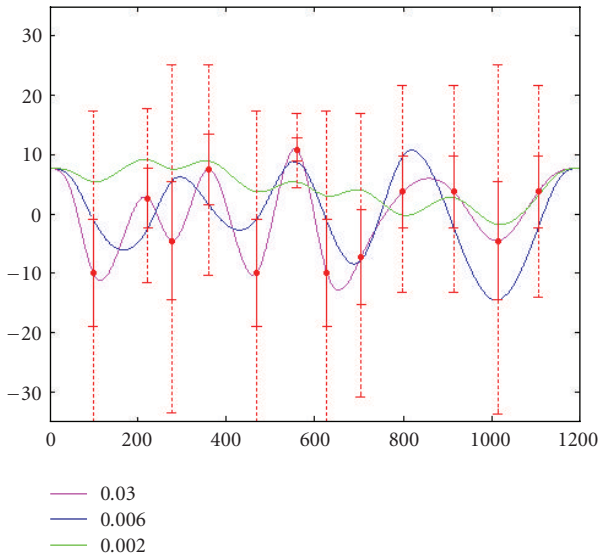


Figure 7: The spline curves for the results shown in Figures 6(a) (pink), 6(b) (blue), and 6(c) (green). The horizontal axis gives time for the speech utterance. The key shows the value of the global acceleration constraint. The red circles are the targets. The solid vertical red bars show the range constraints for Figures 6(a) and 6(b). The dotted bar is the relaxed range constraint for Figure 6(c).

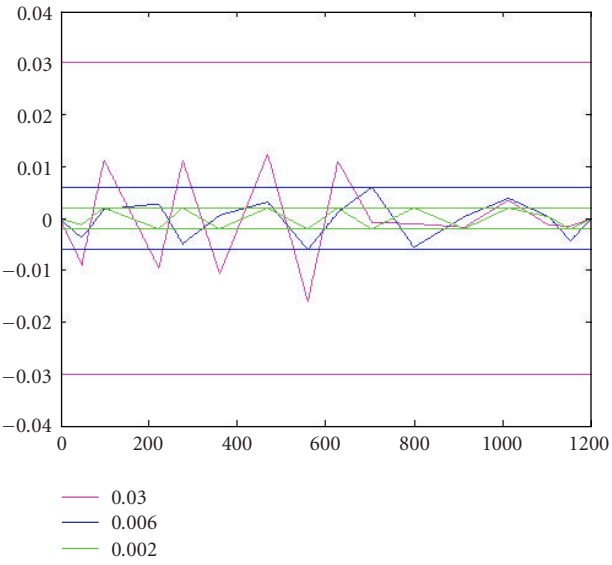


Figure 8: The values of the global acceleration constraints for the results shown in Figures 6(a) (pink), 6(b) (blue), 6(c) (green), and Figure 7. The horizontal axis gives time for the speech utterance. The horizontal lines give the limits of the acceleration constraint in each case.

amount of data used in the optimization approach. However, it is not clear that this produces a suitable set of parameters to control. We are currently considering alternative parameterizations.

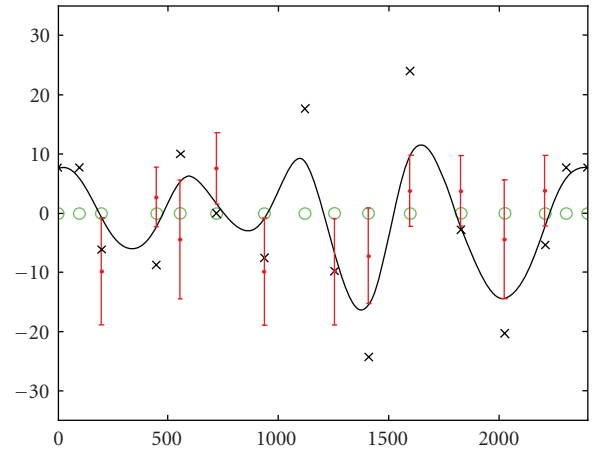


Figure 9: The spline curve for the result shown in Figure 6(d). The global constraint is set to 0.002, and all range constraints are met. The duration of the speech (horizontal axis) is twice as long as Figure 7. The green circles illustrate the knot spacing of the spline, and the x's represent the control points. The solid vertical red bars show the range constraints.

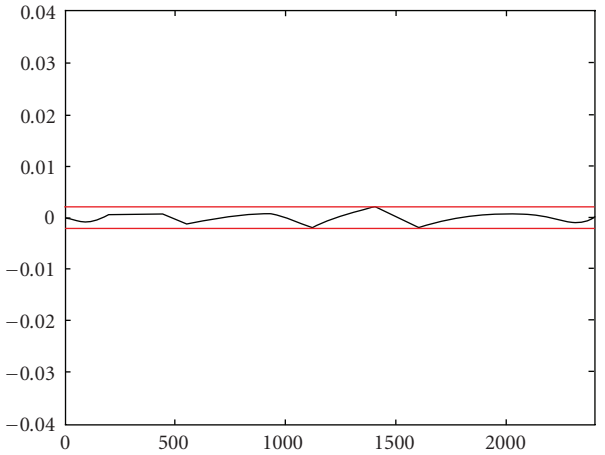


Figure 10: The values of the global acceleration constraint for the result shown in Figure 6(d). The horizontal lines give the limits of the acceleration constraint.

ACKNOWLEDGMENTS

The authors would like to thank Miguel Salas and Jorge Arroyo. They also like to express their thanks to CONACYT.

REFERENCES

[1] F. I. Parke and K. Waters, *Computer Facial Animation*, A K Peters, Wellesley, Mass, USA, 1996.

[2] A. Löfqvist, "Speech as audible gestures," in *Speech Production and Speech Modeling*, W. J. Hardcastle and A. Marchal, Eds., pp. 289–322, Kluwer Academic Press, Dordrecht, The Netherlands, 1990.

[3] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Proceedings of the Computer Animation*, pp. 139–156, Geneva, Switzerland, June 1993.

[4] J. Edge, *Techniques for the synthesis of visual speech*, Ph.D. thesis, University of Sheffield, Sheffield, UK, 2005.
[5] J. Edge and S. Maddock, "Constraint-based synthesis of visual speech," in *Proceedings of the 31st International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '04)*, p. 55, Los Angeles, Calif, USA, August 2004.
[6] A. Witkin and M. Kass, "Spacetime constraints," in *Proceedings of the 15th International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '88)*, pp. 159–168, Atlanta, Ga, USA, August 1988.
[7] O. M. Lazalde, S. Maddock, and M. Meredith, "A Mexican-Spanish talking head," in *Proceedings of the 3rd International Conference on Games Research and Development (CyberGames '07)*, pp. 17–24, Manchester Metropolitan University, UK, September 2007.
[8] P. E. Gill, W. Murray, and M. Wright, *Practical Optimisation*, Academic Press, Boston, Mass, USA, 1981.
[9] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, The MIT Press, Cambridge, Mass, USA, 1998.
[10] B. Dodd and R. Campbell, Eds., *Hearing by Eye: The Psychology of Lipreading*, Lawrence Erlbaum, London, UK, 1987.
[11] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication*, vol. 15, no. 4, pp. 387–421, 2000.
[12] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis System," 2007, http://www.cstr.ed.ac.uk/projects/festival/.
[13] S. Kshirsagar, T. Molet, and N. Magnenat-Thalmann, "Principal components of expressive speech animation," in *Proceedings of the International Conference on Computer Graphics (CGI '01)*, pp. 38–44, Hong Kong, July 2001.
[14] S. Kshirsagar, S. Garchery, G. Sannier, and N. Magnenat-Thalmann, "Synthetic faces: analysis and applications," *International Journal of Imaging Systems and Technology*, vol. 13, no. 1, pp. 65–73, 2003.
[15] S. Kshirsagar and N. Magnenat-Thalmann, "Visyllable based speech animation," *Computer Graphics Forum*, vol. 22, no. 3, pp. 631–639, 2003.
[16] A. P. Benguerel and H. A. Cowan, "Coarticulation of upper lip protrusion in French," *Phonetica*, vol. 30, no. 1, pp. 41–55, 1974.