A Mexican-Spanish Talking Head

Oscar M Martinez Lazalde Steve Maddock Michael Meredith
University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, U.K.
+44(0) 114 2221800
{o.lazalde, s.maddock, m.meredith}@dcs.shef.ac.uk

ABSTRACT
A coarticulation model that overcomes some of the problems of the Dominance functions approach is implemented on a Mexican-Spanish talking head. Some of the important characteristics of this approach are tested and some findings on the way of tuning this approach are mentioned.

Categories and Subject Descriptors
I.3.7 [**Three-Dimensional Graphics and Realism**]: Computer Facial Animation

General Terms
Algorithms, Human Factors.

Keywords
Computer facial animation.

1. INTRODUCTION
There are a number of approaches to producing visual speech and general facial movements, such as pose-based interpolation, concatenation of dynamic units, and physically-based modeling (see [Park96] for a review). We use pose-based interpolation, for which there are two main components. First, a set of static facial postures is created. Second, an interpolation process is defined that will be used to create animation using the static postures. Figure 1 illustrates this. The number of static facial postures needed depends on the range of movement needed in the final facial animation.

For visual speech, the facial postures (visemes) are the shape and position of the articulatory system (lips, teeth/jaw, tongue) at its visual extent for each phoneme in the target language. As an example, the lips would be set in a pouted and rounded position for the /u/ in b**oo**. For English, less than sixty phonemes are needed, but these 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). This means that the technique is low on data requirements, although extra postures are required for further facial postures such as expressions or eyebrow movements for portraying prosody.

The second stage of the pose-based interpolation approach is the interpolation to produce animation. Each of the key poses is described by a set of parameters, which may be as basic as vertex positions in a polygon mesh model, or higher-level parameters describing how to position groups of vertices. Parametric curves can then be fitted through these parameters and used to produce intermediate poses. We refer to the animation path as a trajectory. For visual speech, the speech is broken into a sequence of phonemes (with timing), then these are matched to their equivalent visemes, and then intermediate poses are produced using parametric interpolation. It is this interpolation process which is the key to producing good visual speech.

The shape of the articulatory system is context dependent, an effect known as coarticulation [Lofq90]. 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 [Cohe93]. However, this approach suffers from a number of problems (see [Edge05] for a detailed discussion): only C^0 curve continuity can be asserted, there is no absolute guarantee that a target will be interpolated, higher-level planning is required to control target proximity and context rather than this resulting from a physical dependency, silence is a target that influences surrounding targets, and there are no global parameters to model speaker-independent characteristics. It

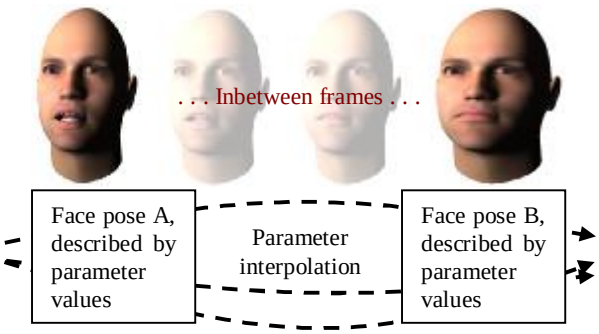


Figure 1. Creating animation frames inbetween given static poses.

addition, a separate parameter tuning process is required for each new visual speaker, and, perhaps more fundamentally, the technique 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 under-articulate, 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 [Edge04; Edge05]), which is based on Witkin and Kass's work on physics-based articulated body motion [Witk88]. Section 2 presents the constraint-based approach.

In Section 3 we demonstrate how the approach is used to create Mexican-Spanish visual speech for a synthetic 3D head. The section outlines the required input data and observations for the constraint-based approach, and shows the results from a synthetic talking head. Finally, section 4 presents conclusions and suggestions for future work.

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 realisation 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 a /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. Given the right values, the global constraint fights with the distribution (or range) constraints to produce a peace where they are both satisfied. Variations can be used to give different trajectories. Extreme values of the global constraint (together with relaxed range constraints) can be used to simulate under-articulation (e.g. mumbling). Ideal targets can be met (e.g. as perhaps used by a stage performer) by relaxing the global constraint. 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 optimisation function is used to create a trajectory that tries to pass as close to the centre of each viseme. We believe this approach better matches the mental and physical activity that produces the coarticulation effect, thus leading to better visual speech. In using a constrained optimisation approach, we need two parts: an objective function, Obj(X) and a set of bounded constraints C_j:

(2.1)

$$\text{minimise} \quad Obj(X)$$
$$\text{subject to} \quad \forall j : \underline{b}_j \le C_j(X) \le \bar{b}_j$$

where $\underline{b}_j$ and $\bar{b}_j$ are the lower and upper bounds. 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.

The following maths is described in detail in [Edge05]. Only a summary is offered here. The particular optimisation function we use is:

(2.2) $$Obj(X) = \sum_i w_i (S(t_i) - V_i)^2$$

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 favoured. Essentially, this governs how much a target dominates its neighbours. 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, e.g. a neutral state such as silence. These are the boundary constraints, as listed in Table 1, which, if necessary, can be used to join trajectories together.

Table 1. Boundary Constraints

Constraints	Action
$S(t_{start}) = \varepsilon_{start}$	Ensures trajectory starts at ε_{start}
$S(t_{end}) = \varepsilon_{end}$	Ensures trajectory ends at ε_{end}
$S(t_{start})' = S(t_{end})' = 0$	Ensures the articulators are stationary at the beginning and end of the trajectory
$S(t_{start})'' = S(t_{end})'' = 0$	Ensures the articulators are in a rest state at the beginning and end of the trajectory

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

(2.3) $$S(t_i) \in [\underline{V}_i, \bar{V}_i]$$

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

If (2.3) and Table 1 are used in Equation (2.2), 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:

(2.4) $$|S(t)''| \le g \qquad \text{where} \quad t \in [t_{start}, t_{end}]$$

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 (2.2) begin to have an effect. The trajectory bends more towards the target where w_i is high relative to its neighbours. Figure 2 gives a conceptual view of this process.

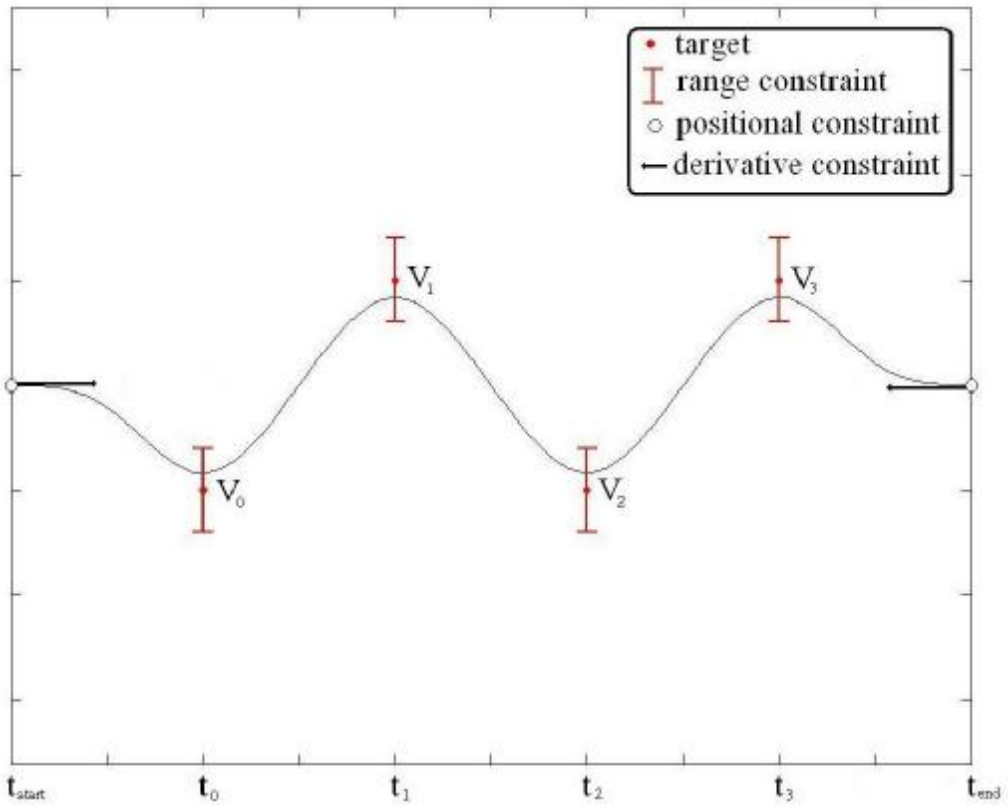


Figure 2. Conceptual view of optimization-based generation of speech trajectories

The speech trajectory S is represented by a cubic non-uniform B-spline. This gives the necessary C^2 continuity to enable (2.4) to be applied.

The optimisation problem is solved using a variant of the Sequential Quadratic Programming (SQP) method as it is proposed in [Witk88]. The SQP algorithm requires the objective function described in (2.2). 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 (2.5), (2.6) and (2.7). The iterative process finishes when an optimisation criterion is met (discussed in section 5).

$$\Delta X_{obj} = -H_{obj}^{-1} \begin{pmatrix} \frac{\partial Obj}{\partial X_1} \\ \vdots \\ \frac{\partial Obj}{\partial X_n} \end{pmatrix} \tag{2.5}$$

$$\Delta X_{cstr} = -J_{cstr}^{+}(J_{cstr}\Delta X_{obj} + C) \tag{2.6}$$

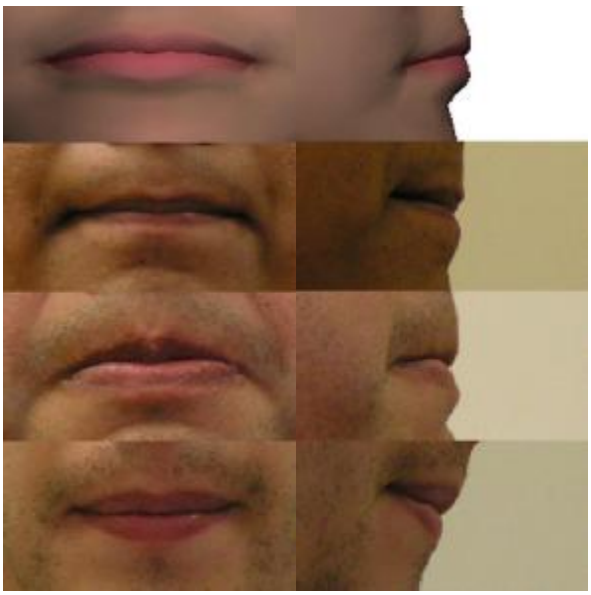
$$X_{j+1} = X_j + (\Delta X_{obj} + \Delta X_{cstr}) \tag{2.7}$$

3. INPUT DATA FOR THE RANGE CONSTRAINTS

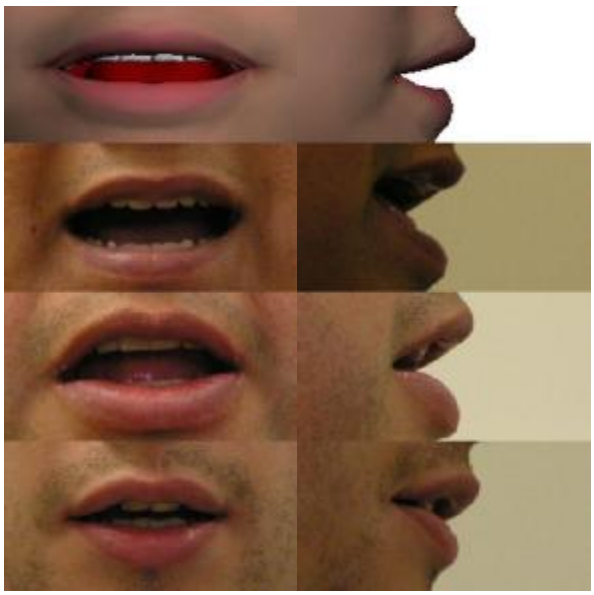
Observation of real Mexican-Spanish speakers was done to give values for the Range constraints. They were asked to make each of the viseme shapes and they were photographed from front and side views. The visemes are defined as shown in Table 2.

In Figure 3 it can be observed the front and side views of the 3D model and three people doing a consonant viseme M and a vowel viseme A. It can be observed differences on the shape of the lips between speakers. These differences can be due to physical differences and/or due to different manners of articulation. The differences in manner can be encoded in the range constraints. The same can be observed for the vowel A in figure 3b. The 3D model visemes were done by using the software Facegen.

Even observing the same speaker there will be differences on the manner of articulation of a phoneme. Figure 4 illustrate the shapes of the mouth of a speaker articulating the word “ama”, in the top row is shown how the speaker articulates in a normal way, in the middle row is over articulating and in the bottom row is mumbling. After observing the differences is possible to define upper and lower values for the range constraints, as said before, these values could change between speakers.



a) Front and side view of the viseme M



b) Front and side view of the viseme A

Figure 3

There is another fact that affects the manner of articulation of a phoneme, as mentioned before, it is the coarticulation. Depending on the context of the phoneme it will be articulated in a different way, this can be observed in figure 5. The speaker was recorded pronouncing the words “ama”, “eme” and “omo” and the frames containing the center of the phoneme m where extracted. It can be observed that the shape of the mouth is more rounded in the middle row than in the other rows due that the phoneme *m* is surrounded by the rounded vowel *o*.

Table 2. Mexican-Spanish viseme definition

Phoneme	**Viseme name**
silence,h	NEUTRAL
j,g	J
b,m,p,v	B_M_P
a	A
ch,ll,y,x	CH_Y
d,s,t,z	D_S_T
e	E
f	F
i	I
c,k,q	K
n,ñ	N
o,u	O
r	R

l	L
w,gu	W

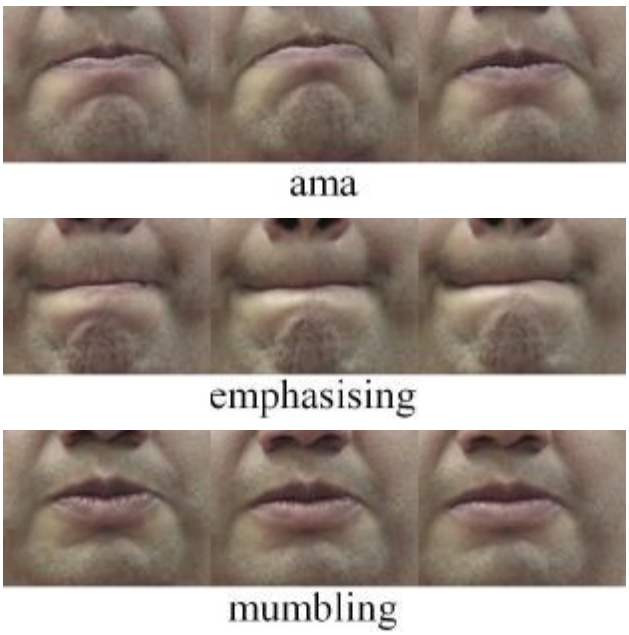


Figure 4. Manner differences of the phonene *m* in the word "ama"

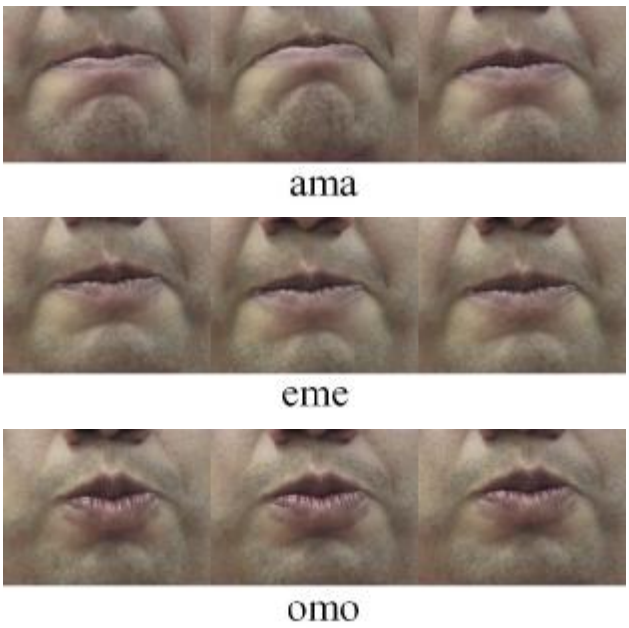


Figure 5. Manner differences of the phonene *m* due to coarticulation

There are other factors that could affect the manner of articulation of a phoneme such as mood. The covering of such situations will be done in future research.

4. RESULTS

A talking head was implemented. It has a main C++ module which is in charge of communicating the rest of the modules (see figure 6). This module gets texts as input, gets the phonetic transcription, audio wave and timing from a Festival server, gets the viseme data according to the phonetic transcription, defines the optimization problem and passes it to a MATLAB routine which contains the SQP implementation, a spline definition is returned and then it generates the rendering of the 3D face in synchronization with the audio wave.

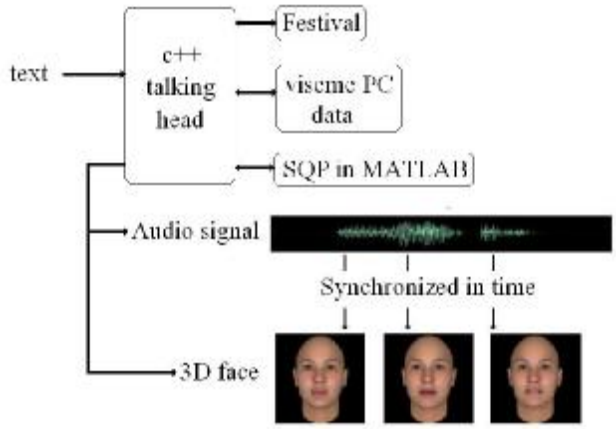


Figure 6. Talking head system

The 3D mesh of each viseme consist of 1504 vertices which is a considerable amount of data if we take in count that for each vertex the optimization process has to be done. To overcome this problem Principal Component Analysis (PCA) is used to reduce the amount of data.

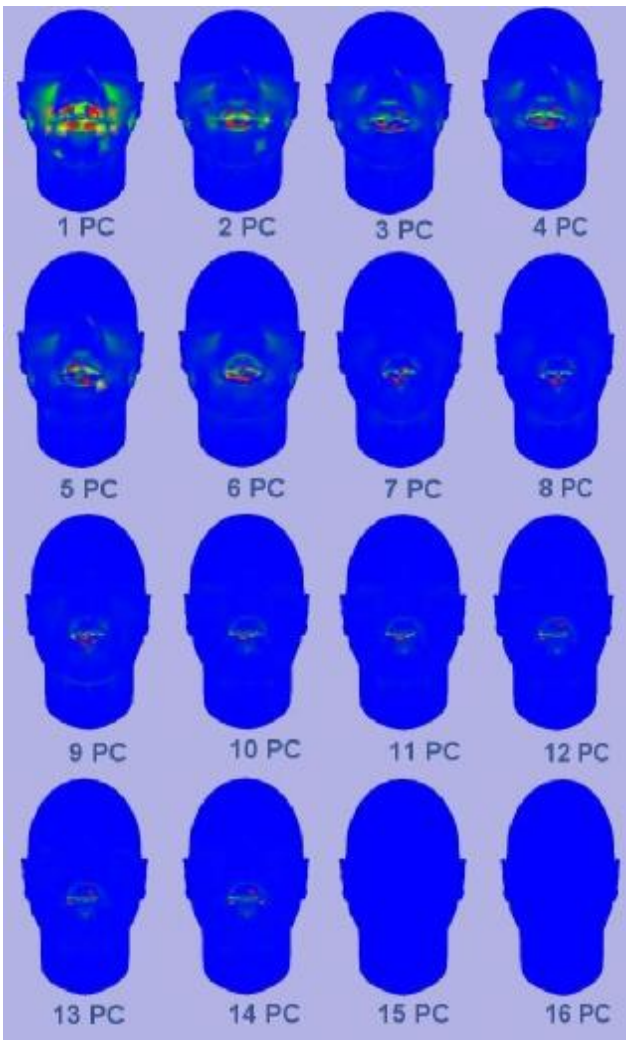


Figure 7. Differences between original mesh and mesh reconstruction using different number of PCs

PCA allows representing each of the visemes by a vector of weight values. This technique lets reconstruct a vector v_i that belongs to a randomly sampled vector population V using (4.1).

$$V = \{v_0, v_1, \mathbf{L}, v_s\}$$

$$v_i = u_V + \sum_{j=1}^{s} e_j b_j \qquad 0 \le i \le s \tag{4.1}$$

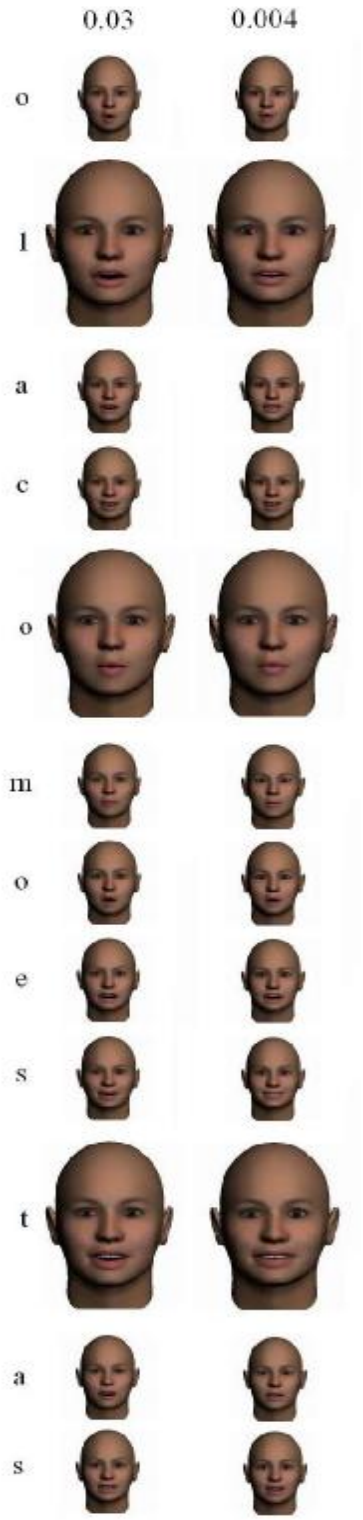


Figure 8. Face positions of the Spanish sentence "hola, ¿cómo estas?". Left column: meeting targets (global constraint 0.03). Right column: targets not met (global constraint 0.004).

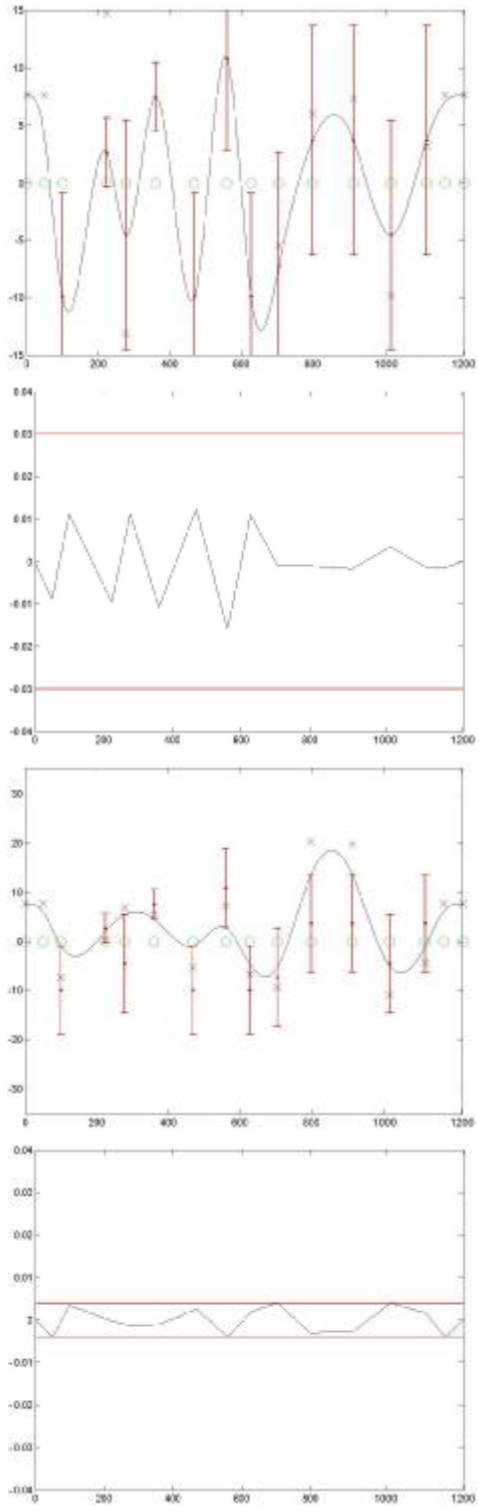


Figure 9. First two rows: spline curve and global acceleration when all the targets are met. Last two rows: spline curve and global acceleration when global constraint is reduced (red line), targets are not met.

Where u_V is the mean vector; e_i are the eigenvectors obtained after applying the PCA technique and b_i are the weight values. With this technique, it is possible to reconstruct at the cost of a minimal error any of the vectors in the population using a reduced number of eigenvectors e_i and its corresponding weights b_i. To do the reconstruction all the vectors share the reduced set of eigenvectors e_i (PCs) but they use different weights b_i for each of those eigenvectors.

With this technique, the calculation for 1504 vertices is reduced to do calculations for only 8 weights. 8 PCs were chosen by observing the differences between the original mesh and the reconstructed mesh using different number or PCs, this can be observed in figure 7. 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 [Edge05], and Kshirsagar et al have used 7 [Kshi01], 8 [Kshi03a] and 9 [Kshi03b].

It is the PCs that are the parameters (targets) that need to be interpolated. The range constraints for the constraint-based approach described in section 2 need to be defined in terms of a range for each PC. The acceleration constraint is for each PC. The ranges were defined by comparing against the real faces.

The talking head was tested with the sentence "hola, ¿cómo estas?". In figure 8 the resulting 3D faces for two configurations are shown, the left column shows the results of the animation with a global constraint with value 0.03, the right column shows the result of the animation with a global constraint with value 0.004. Comparing figure 8 left (global constraint equal to 0.03) against figure 8 right (global constraint equal to 0.004) it can be observed differences in the mouth opening. 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 figure 9 (first two rows), it can be observed that all the targets are met for the 1st PC, the green points represent the knots of the spline, the x represent the control points, the red points represent the targets and the red lines represent the range of each target. The global constraint value doesn't influence the result. The global acceleration was restricted to 0.004 to not reach some targets. This can be observed in figure 9 (last two rows). It can be observed that now the acceleration is restricted by the global acceleration constraint (red line), this causes some targets not to be met as the spline curve indicates. It can be observed that changing the value of the global constraint will lead to different animation curves. Here we have to point that in all the animation curves, both, the global and the range constraint were coexisting, but making the global constraint smaller could lead to an unstable system where both kinds of constraints will fighting. To make the system stable 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 are interested preserving a speaker dependant animation we will finish relaxing the global constraints as the range constraints will encode the boundaries of the manners of articulation of that speaker. If we are interested on making mumbling effects or making animations that were we are not interested in preserving the speaker manners of articulation then the range constraint will have to be relaxed.

5. CONCLUSIONS

The variability due to different speakers, due to the context and due to mood and tiredness is encoded within the constraint-based approach.

Variations of the global acceleration constraint can ensure different trajectories which make this approach suitable to reproduce variations in manner due emphasizing or mumbling.

Looking again at figure 5 and observing at the shape that the viseme M gets in the middle of the sentence *omo*, make us wonder about the suitability of using PCs along with this technique, as the first PC is related with mouth closure it is not enough tuning the ranges to get the rounding shape, doing some experiments was found that the rounding is a the result of a combination of the rest of the PCs since the variables are not independent. We plan to do more experiments on alternative sets of parameters to describe mouth shape and understanding more what each PC encode.

Future work include measuring range constraints for static visemes using continuous speaker video. This includes lip tracking and extraction of still not defined mouth characteristics. The tuned system will evaluated against original continuous real visual speech.

6. ACKNOWLEDGMENTS

Thanks to Miguel and Jorge.

7. REFERENCES

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

[2] Edge, J. *Techniques for the Synthesis of Visual Speech*. Ph.D. Thesis, University of Sheffield, Sheffield, England, 2005.

[3] Edge, J. and Maddock, S. (2004). Constraint-based synthesis of visual speech. In *Proceedings of SIGGRAPH'04 Sketches Programme.*

[4] Kshirsagar, S., Molet, T., et al. (2001). Principal Components of Expressive Speech Animation. *In: International Conference on Computer Graphics, 2001*. IEEE Computer Society, 38-44.

[5] Kshirsagar, S., Garchery, S., et al. (2003a). Synthetic faces: Analysis and applications. *International Journal of Imaging Systems and Technology*, **13**(1): 65-73.

[6] Kshirsagar, S. and Magnenat-Thalmann, N. (2003b). Visyllable Based Speech Animation. *Computer Graphics Forum*, **22**(3): 631.

[7] Lofqvist, A. Speech as audible gestures. *In Speech Production and Speech Modeling* (Eds, Hardcastle, W. J. and Marchal, A.), Kluwer, 289-322.

[8] Parke, F.I. and K.Waters. *Computer Facial Animation*. AK Peters, 1996

[9] Witkin, A. and Kass, M. Spacetime constraints. In *Proceedings of SIGGRAPH 1988*, 159-168.