Spacetime Constraints for Viseme-based Synthesis of Visual Speech

James D. Edge and Steve Maddock

Technical Report CS-04-03
Department of Computer Science
University of Sheffield
UK

1 Introduction

This technical report concerns the animation of facial movement during speech production. In this work we consider speech gestures as trajectories through a space containing all visible vocal tract postures. Within this *visible speech space*, visual-phonemes (or visemes) are defined as collections of vocal tract postures which produce similar speech sounds (i.e. an individual phoneme in audible speech). Furthermore, visemes discount vocal tract information regarding non-visible articulations, i.e. /p,b,m/ can be considered an individual phoneme because the variation whilst audibly distinct is invisible. This definition contrasts with many techniques in which the terms viseme and morph-target could be used interchangably (e.g. [Cohen and Massaro 1993]). A speech trajectory will always interpolate the visemes corresponding to its phonetic structure (i.e. there is a direct mapping from audio → visual speech). However, as visemes are not individual targets we must determine how the trajectory passes through each of the viseme clusters according to both physical constraints and context; this is the notion of coarticulation [Öhman 1967; Löfqvist 1990].

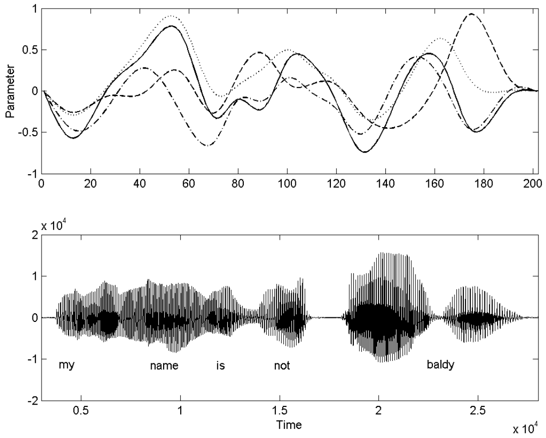


Figure 1: Several parameter trajectories and speech waveform generated for the utterance 'my name is *not* baldy'.

2 Method

Our system works by generating trajectories which pass through appropriate visemes as specified by the phonetic structure of the target utterance. Each viseme, V, is regarded as normally distributed; that is the ideal vocal tract configuration is located at the mean, μ_V, and the scale of allowable variation from this ideal is defined by the standard deviation, σ_V, from that mean. The deformability of a viseme in context is directly correlated to σ_V; that is highly deformable visemes (e.g. mouth opening in /t/) will exhibit large σ_V and, conversely, non-deformable visemes (e.g. lip contact in /p,b,m/) will only exhibit small variations. A visual speech utterance is described by a sequence of viseme-time pairs, e.g. the word 'cat' corresponds to the sequence $\left[\{/k/,t_0\},\{/ae/,t_1\},\{/t/,t_2\}\right]$

In order to generate viseme transitions for a given speech utterance we apply a technique similar to the spacetime constraints method used for articulated body animation [Witkin and Kass 1988]. The use of constrained optimization techniques for facial animation require us to define both an objective function, R, defining the goodness of any step in the optimization, and a number of constraints, C_j, whose bounds ($\underline{b}_j$ and $\bar{b}_j$) maintain the physical properties of speech movement. The optimization procedure determines the speech trajectory, S, for which $R(S)$ is optimal; i.e. finding the solution to (1).

$$\textbf{min } R(S) = \sum_i \omega_{Vi} \| S(t_i) - \mu_{Vi} \|^2 \qquad (1)$$

$$\textbf{subject to } \underline{b}_j \leq C_j \leq \bar{b}_j$$

In (1), the objective function optimizes the distance between the speech trajectory and each of the *ideal* viseme centres, μ_{Vi}, at the appropriate instances in time, t_i. This implies that in the absence of any physical constraints upon the speech trajectory that all the relevant μ_{Vi} would be interpolated. However, this is not the case in natural speech movements where visemes are met to a variable degree depending upon their importance/dominance over the speech utterance (i.e. due to coarticulation). For this reason the targets are weighted in accordance with their dominance using the ω_{Vi} weights. The dominance of each viseme lies in the interval $\omega_{Vi} \in [0,1]$, and in practical terms these weights will vary with each parameter representing our visible speech space.

A direct interpolation is inadequate to represent speech trajectories, and so our constraints must reflect the fact that parameters controlling the vocal tract can only change at a given rate. These constraints are necessarily specific to the parameterization of the individual model we are controlling. For a physical model of facial expression (e.g. [Lee et al. 1995]), the forces applied by muscles can be directly constrained so that they conform to the onset/offset characteristics reported in [Fung 1993]. For geometric models of facial expression, parametric acceleration across the utterance can be constrained to similar effect. These constraints are applied to ensure that the speech trajectory *does not* interpolate the viseme centres, but instead is varyingly attracted according to the dominance of the respective viseme.

The method we use differs from most spacetime methods in that we do not aim to interpolate a set of key frames, nor do we optimize any form of energy conservation term. This is because in natural speech we will never actually meet our ideal targets (the μ_V), and for this reason there ought not to be any slack in the speech trajectory to remove. This also contrasts with methods such as [Ruttkay 1999] which use constraints to augment keyframe approaches.

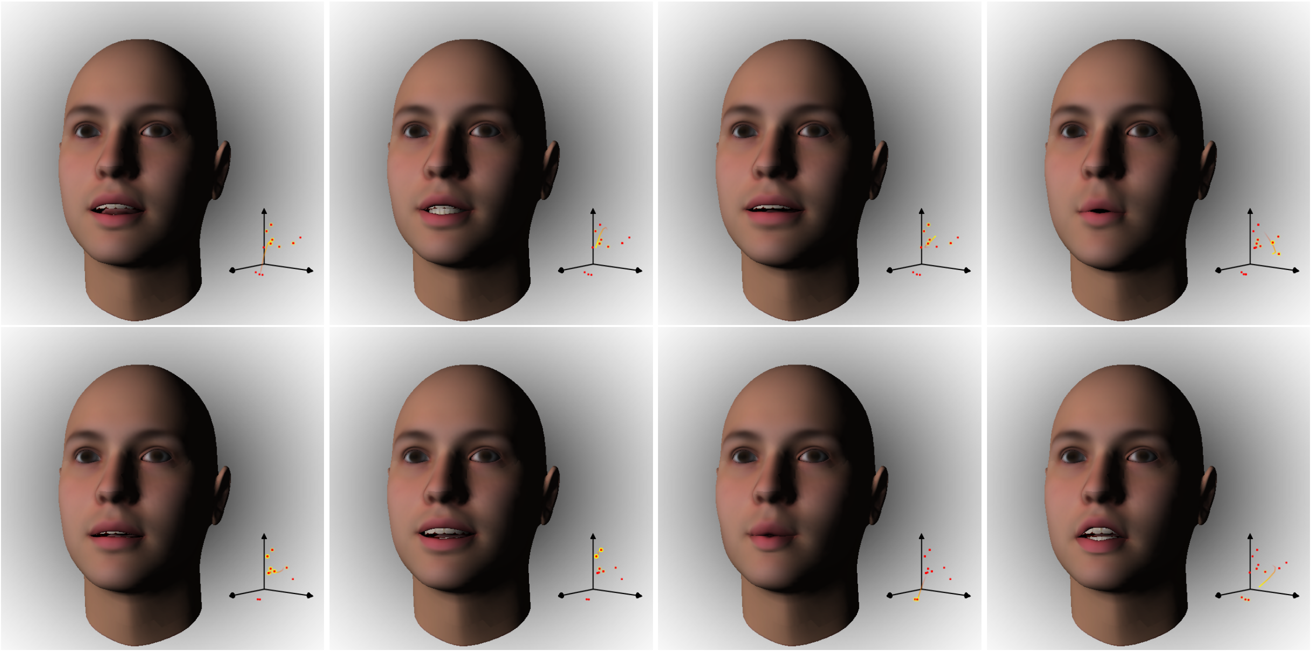


Figure 2: Frames from an animation including local trajectories through the first three principal components of viseme space.

2.1 Results

Figures 3 and 4 demonstrate the effect of varying the global acceleration constraint and parameters for viseme dominance for simple speech trajectories. These two types of parameter in conjunction can be used to simulate simple speech coarticulation for visual speech. The global constraint, which for these example trajectories simply limits the parametric acceleration, dampens the possible motion thereby preventing the μ_{Vi} from being met. In Figures 3 (a) and 4 (a) we can see the dampening affect of increasing this constraint; conversely relaxing the constraint will increasingly allow the trajectory to better meet the targets until they are exactly interpolated.

The weights, ω_{Vi}, control the dominance that an individual viseme exerts over the speech trajectory. In Figures 3 (b) and 4 (b) the effect of varying the dominance of an individual viseme is demonstrated. Lower values for ω_{Vi} will lead to V_i having less effect over the trajectory, and conversely high values will lead to greater influence. For the case where each of the visemes has equal dominance the trajectory will tend to an average, equally meeting all of the targets.

Figure 2 shows frames from an animation generated using the described method. Further animations can be found at: `http://www.dcs.shef.ac.uk/~jedge/`.

2.2 Conclusions and Future Work

The combination of global constraints and dominance weights allow simple speech trajectories, such as those in Figure 1, to be generated. The model of coarticulation as described in this paper necessarily makes some naïve assumptions, in particular our model assumes symmetry in the dominance of an individual viseme over an utterance. This is not the case in natural speech where forward (preparatory) and backward (carry-over) coarticulation have been observed. However, the power of constrained optimization techniques lies in their extensibility, and such assymetries could be accounted for by restricting motions such that they follow the onset/offset patterns of muscular contraction [Fung 1993].

Further constraints can be added both to speed up the solution of the system and to mimic the physical properties of speech production. For example, we apply constraints so that $S(t_i)$ lies within $\pm 3\sigma_{Vi}$ of μ_{Vi}; this can be seen as a minimum qualification for the speech trajectory to be producing the appropriate speech audio. An example of a speech-oriented constraint would be enforcing that the lips are moving apart at the audio centre of a bilabial plosive (e.g. ***p****it* or ***b****ead*). Because we can arbitrarily add further constraints the model can be iteratively refined to get as close as possible to real speech. For models such as [Cohen and Massaro 1993] this is not possible, and shortcomings in the method prevent certain forms of articulation from being accurately reproduced (see [Goff and Benoît 1996]). The method described here is seen as an early iteration in modelling coarticulation with greater flexibility in representing observed speech characteristics.

References

COHEN, M., AND MASSARO, D. 1993. Modeling coarticulation in synthetic visual speech. *Computer Animation '93*.

FUNG, Y. C. 1993. *Biomechanics - Mechanical Properties of Living Tissues*, second ed. Springer-Verlag.

GOFF, B. L., AND BENOÎT, C. 1996. A text-to-audiovisual-speech synthesizer for french. In *Proceedings ICSLP'96*, vol. 4, 2163–2166.

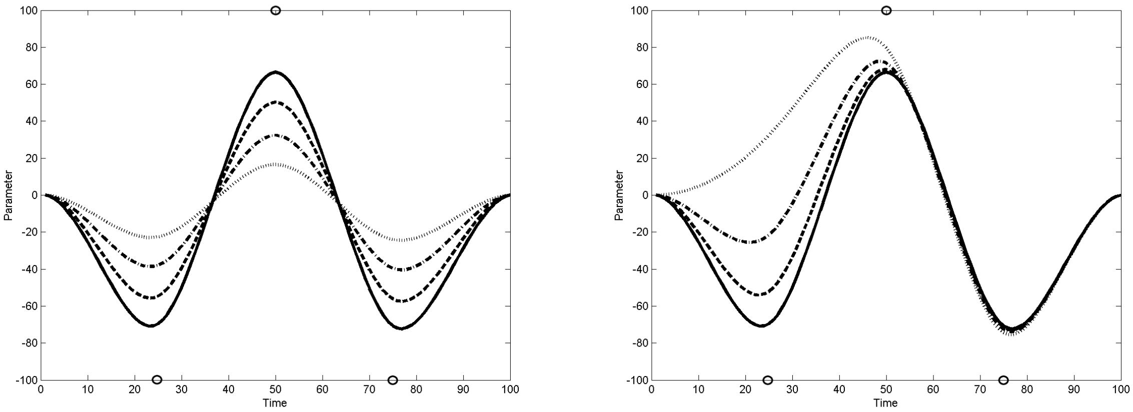
LEE, Y., TERZOPOULOS, D., AND WATERS, K. 1995. Realistic modeling for facial animation. *Computer Graphics 29*, Annual Conference Series, 55–62.

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

ÖHMAN, S. 1967. Numerical model of coarticulation. *Journal of the Acoustical Society of America 41*, 310–320.

RUTTKAY, Z. 1999. Constraint-based facial animation. Tech. Rep. INS-R9907, Centrum voor Wiskunde en Informatica (CWI).

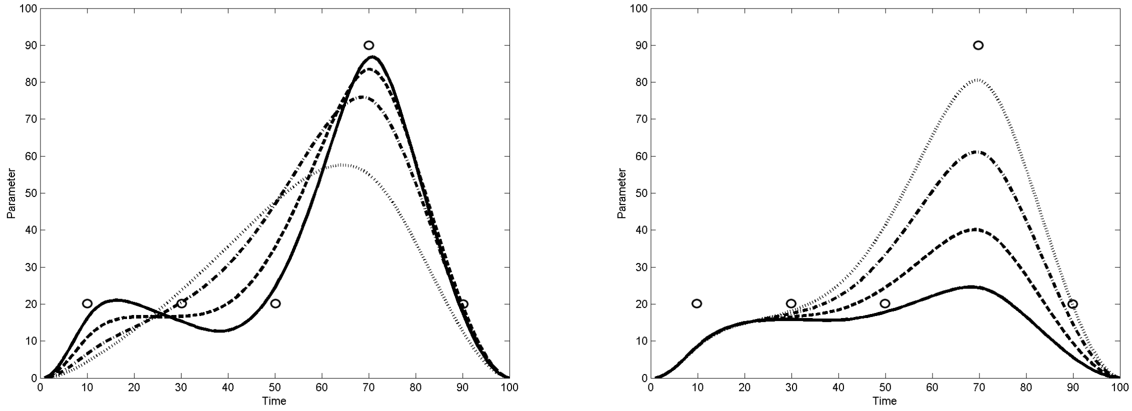
WITKIN, A., AND KASS, M. 1988. Spacetime constraints. In *Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, 159–168.



(a)

(b)

Figure 3: Varying parameters for a simple trajectory (three visemes: $\{\mu_0 = -100, \omega_0 = 1.0, t_0 = 25\}, \{\mu_1 = 100, \omega_1 = 1.0, t_1 = 50\}, \{\mu_2 = -100, \omega_2 = 1.0, t_2 = 75\}$). (a) demonstrates the effect of increasingly constraining parametric acceleration on the final speech trajectory (the dotted trajectory is most constrained, solid is least); this effectively dampens the motion preventing all targets from being met. (b) demonstrates the effect of varying the dominance of the first viseme from no dominance ($\omega_0 = 0.0$, dotted), to equal dominance ($\omega_0 = 1.0$, solid). A combination of these two types of variable can be used to model basic speech coarticulation.



(a)

(b)

Figure 4: Further examples of our coarticulation model for a slightly more complex trajectory ($\{\mu_0 = 20, \omega_0 = 0.6, t_0 = 10\}, \{\mu_1 = 20, \omega_1 = 0.1, t_1 = 30\}, \{\mu_2 = 20, \omega_2 = 0.05, t_2 = 50\}, \{\mu_3 = 90, \omega_3 = 1.0, t_3 = 70\}, \{\mu_4 = 20, \omega_4 = 0.6, t_4 = 90\}$). (a) decreasing the global constraint allows all the targets to be better met. (b) decreasing the dominance of the fourth segment reduces its affect over the trajectory (notice that this affect is exerted over several surrounding visemes).