Constraint-based Synthesis of Visual Speech

James D. Edge*and Steve Maddock†
Department of Computer Science, University of Sheffield, UK

1 Introduction

This sketch 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). This definition is distinct from 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 visemes according to both physical constraints and context; this is the notion of coarticulation [Löfqvist 1990].

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 will exhibit large σ_V and conversely non-deformable visemes 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 $[\{/k/,t_0\},\{/ae/,t_1\},\{/t/,t_2\}]$.

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, which maintain the physical properties of speech movement. The optimization procedure determines the speech trajectory, S, for which $R(S)$ is optimal subject to $\forall_j : min_j \leq C_j(S) \leq max_j$. The method we use varies 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.

The main assumption in our system is that with no physical or contextual constraints each of the viseme centres, μ_V, would be directly interpolated. This implies that our objective function is $R(S) = \sum_i \omega_{Vi} \|S(t_i) - \mu_{Vi}\|^2$, that is optimizing the weighted square distance to the ideal vocal tract shape for each phoneme in an utterance. The weight, $\omega_{Vi} \in [0,1]$, represents the fact that not all visemes will have an equal dominance over an utterance, which in the absence of constraints will have no effect upon the result. In practice ω_{Vi} will vary with each parameter representing our viseme.

*e-mail: j.edge@dcs.shef.ac.uk
†email: s.maddock@dcs.shef.ac.uk

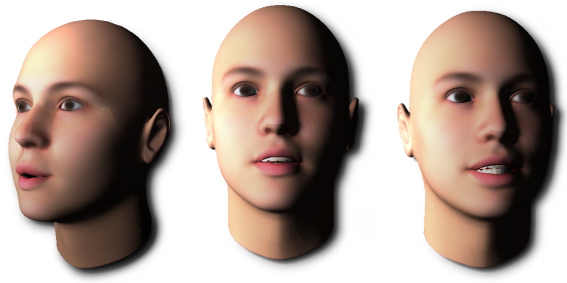

Figure 1: Frames from a speech animation.

A direct interpolation is inadequate to represent speech trajectories, and so our constraints must reflect the fact that the 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.

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 is the 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*).

The power of this technique, as opposed to models such as [Cohen and Massaro 1993], lies particularly in its extensibility. Previous methods make no assertion as to the physical properties of the motion, and so capturing the nature of speech movements is a laborious matter of refining parameter sets. In the described model we can explicitly define constraints for the resulting trajectory and thus any rule enforcable upon a curve can be directly applied. (example animations can be found at: `http://www.dcs.shef.ac.uk/~jedge/`)

References

COHEN, M., AND MASSARO, D. 1993. Modeling coarticulation in synthetic visual speech. *Computer Animation '93*, 131–156.

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

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.

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