Synthetic Animation of Deaf Signing Gestures

Richard Kennaway

School of Information Systems, University of East Anglia**

Abstract

We describe a method for synthesizing deaf signing animations from a high-level description of signs in terms of the HamNoSys transcription system.

1 Introduction

1.1 Background

The object of the ViSiCAST project is to facilitate access by deaf citizens to information and services expressed in their preferred medium of sign language. ViSiCAST aims to provide support in three distinct application areas: broadcasting, face-to-face transactions, and the World-Wide Web (WWW). A central feature of the project is its use of computer-generated virtual humans, or avatars, to present deaf signing; hence, the technical activity of the project focuses on two areas: language processing technology and avatar technology. For an introductory account of the whole project the reader is referred to [2].

In outline, the task of signing textual content is decomposed into the following sequence of transformations:

1. from text to semantic representation;
2. from semantic representation to a morphological representation, which latter is sign-language specific;
3. from morphology to a signing gesture notation;
4. from signing gesture notation to avatar animation.

This paper deals with the last of these steps, and describes an initial attempt to create synthetic animations from a gesture notation, to supplement or replace our current use of motion capture.

** School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, U.K. Email: jrk@sys.uea.ac.uk

* We acknowledge funding from the European Union under the Framework V IST Programme (Grant IST-1999-10500).

1.2 Motion capture vs. synthetic animation

ViSiCAST has developed from two previous projects, one in broadcasting, *Sign-Anim* (also known as *Simon-the-Signer*) [8, 9, 14], and one in Post Office transactions, *Tessa* [7]. Both these applications use a signing avatar system based on *motion capture*: before a text can be signed by the avatar, the appropriate lexicon of signs must be constructed in advance, each sign being represented by a data file recording motion parameters for the body, arms, hands, and face of a real human signer. For a given text, the corresponding sequence of such motion data files can be used to animate the skeleton of the computer-generated avatar. The great merit of this system is its almost uncanny authenticity: even when captured motion data is "played back" through an avatar with physical characteristics very different from those of the original human signer, the original signer (if already known to the audience) is nevertheless immediately recognizable in the result.

On the other hand, motion capture is not without drawbacks.

- There is a substantial amount of work involved in setting up and calibrating the equipment, and in recording the large number of signs required for a complete lexicon.
- It is a non-trivial task to modify captured motions.

There are several ways in which we would wish to modify the raw motion capture data. Data captured from one signer might be played back through an avatar of different body proportions. One might wish to change the point in signing space at which a sign is performed, rather than recording a separate version for every place at which it might be performed. Signs recorded separately, and perhaps by different signers, need to be blended into a continuous animation. Much research exists on algorithmically modifying captured data, though to our knowledge none is concerned specifically with signing, a typical application being modification of a walking character's gait to conform to an uneven terrain. An example is Witkin and Popović's "motion warping" [10, 15]. We are therefore also interested in synthetic animation: generation of the required movements of an avatar from a more abstract description of the gestures that it is to perform, together with the geometry of the avatar in use. The animation must be feasible to generate in real time, within the processing power available in the computers or set-top boxes that will display signing, and the transmitted data must fit within the available bandwidth.

Traditional (i.e. non-computer) animation is a highly laborious process, to which computers were first introduced to perform in-betweening, creating all the frames between the hand-drawn keyframes. Even when animations are entirely produced on computer, keyframes are still typically designed by hand, but using 3D modelling software to construct and pose the characters. In recent years, physics-based modelling has come into use, primarily for animating inanimate objects such as stacks of boxes acted on by gravity. Synthetic animation of living organisms poses a further set of problems, as it must deal not only with gravity and the forces exerted by muscles, but also with the biological control systems

that operate the muscles. It is only in the last few years that implementation techniques and fast hardware have begun to make synthetic real-time animation of living movement possible. The tutorial courses [3, 4] give an overview of this history and the current state of the art.

We present here a simplified biomechanical model dealing with the particular application of signing and the constraints of real-time animation.

2 Description of signs

We start from the HamNoSys [12] notation for transcribing signing gestures developed by our partners at IDGS, University of Hamburg. For use in the ViSiCAST project, we have developed a version of HamNoSys encoded in XML [1], called SiGML (Signing Gesture Markup Language). This is an alternative syntactic representation to facilitate computer processing. As HamNoSys has been designed to be read by humans rather than computers, we use the former for the presentation of examples in this paper.

HamNoSys breaks each sign down into components such as hand position, hand orientation, hand shape, motion, etc. The current version only records the manual components of signs; the recording of facial components is a subject of current research by its developers. When a notation for facial components is available we intend to extend the work reported here to include them. We will not attempt to describe HamNoSys in detail (see the cited manual), but indicate its main features and the issues they raise for synthetic animation.

A typical HamNoSys transcription of a single sign is displayed in Figure 1. This is the DGS (German Sign Language) sign for "GOING-TO". The colon sign

Fig. 1. HamNoSys transcription of the DGS sign for GOING-TO

specifies that the two hands mirror each other. The next three glyphs specify the starting position: the finger and thumb are extended with the other fingers curled, the index finger pointing up and forwards, and the palm of the right hand facing to the left. The part in square brackets indicates the motion: forwards, in an arc curved in the vertical place, while changing the orientation of the hand so that the index fingers point forwards and down. HamNoSys describes the physical action required to produce the sign, not the sign's meaning.

Note that HamNoSys describes signs in terms of basic concepts which are not themselves given a precise meaning: "close to", "chest level", "fast", "slow", etc. Other aspects are not recorded at all, and are assumed to take on some "default" value. For example, HamNoSys records the positions of the hands, but not of the rest of the arms. Shoulder shrugs or raising of the elbows can be notated, but

for signs which do not require such movements, the positions of shoulders and elbows are omitted, and are assumed to be in an ordinarily relaxed position. This is deliberate: only those parts of the action are transcribed which are required to correctly form the sign, and only with enough precision as is necessary. People learning to sign learn from example which parts of the action are significant, and how much precision is required for good signing. To synthesise an animation from a HamNoSys description requires these choices to be made by the animation algorithms.

3 Synthesis of static gestural elements

3.1 Disambiguation

We illustrate by an example how we have approached the task of making precise the fuzzy definitions of HamNoSys components.

HamNoSys defines a set of 60 positions in the space in front of the signer. These are arranged in a grid of four levels from top to bottom, five from left to right, and three from near to far. The four vertical levels are indicated by the glyphs (shoulder level), (chest level), (abdomen level), and (below abdomen level). For any given avatar, we define these to be respectively at heights s, $(s+e)/2$, e, and $(e+w)/2$, where s, e, and w are the heights of the shoulder, elbow, and wrist joints of the standing avatar when both arms hang vertically. We have defined these heights in terms of the arms rather than the torso, because these measurements are guaranteed to be present for any avatar used for signing.

HamNoSys indicates left-to-right location by a modification to these glyphs: the five locations at chest level are represented by the glyphs: , , , , and . We define their left-to-right coordinates in terms of the positions of the shoulders: centre is midway between the shoulders, and the points left and right are regularly spaced with gaps of 0.4 times the distance between the shoulders.

The three distances from the avatar are notated in HamNoSys by)((near), no explicit notation for neutral, and ↝ (far); we defined these in terms of the shoulder coordinates and the length of the forearm.

An important feature of this method of determining numerical values is that it is done in terms of measurements of the avatar's body, and can be applied automatically to any humanoid avatar. The precise choice of coordinates for these and other points must be judged by the quality of the signing that results. The animation system is currently still under development, and we have not yet brought it to the point of testing.

Hand orientations are described by HamNoSys in terms which are already precise: the possibilities are all of the 26 non-zero vectors (a, b, c), where each of a, b, and c is -1, 0, or 1. When more precision is required, HamNoSys also allows the representation of any direction midway between two such vectors. These directions are the directions the fingers point (or would point if the fingers were extended straight); the orientation of the palm around this axis takes 8 possible values at 45 degree intervals.

To complete the specification of HamNoSys positions requires definitions along similar lines of all the positions that are significant to HamNoSys. In addition to these points in space, HamNoSys defines many contact points on the body, such as positions at, above, below, left, or right of each facial element (eyes, nose, cheeks, etc.), and several positions on each finger and along the arms and torso. The total number of positions nameable in HamNoSys comes to some hundreds. Any avatar for use in signing must include, as part of its definition, not only the geometry of its skeleton and surface, and its visual appearance, but also the locations of all of these "significant sites".

3.2 Inverse Kinematics

Given a definition of the numerical coordinates of all the hand positions described by HamNoSys, we must determine angles of the arm joints which will place the hand in the desired position and orientation. This is a problem in "inverse kinematics" (forward kinematics being the opposite and easier problem, of computing hand position and orientation from the arm joint angles).

The problem can mostly be solved by direct application of trigonometric equations. Two factors complicate the computation: firstly, the arm joint angles are not fully determined by the hand, and secondly, care must be taken to ensure that physiologically impossible solutions are avoided.

If the shoulder joint remains fixed in space, and the hand position and orientation are known, then one degree of freedom remains undetermined: the arm can be rotated about the line from the shoulder to the wrist joint. If the sign being performed requires the elbow to be unusually elevated, this will be notated in the HamNoSys description; otherwise, a choice must be made by the animator as to what is a natural position for the elbow. The first solution we adopted required the elbow to lie vertically below the line from shoulder to wrist. This gives satisfactory results for placements of the hand in the same half of signing space as the shoulder, but reaches across the body resulted in the upper arm penetrating the torso. A correction was therefore made to rotate the elbow away from the body when necessary to maintain a certain minimum separation. In addition, for such reaches, and for reaches into the "far" part of signing space, greater realism is obtained by using the sternoclavicular joint to let the shoulder move some distance towards the target point.

For positions around the head, further care must be taken to avoid the hand penetrating the head. It should be noted that HamNoSys itself does not attempt to syntactically exclude the description of physiologically impossible signs. One can, for example, specify a hand position midway between the ears. This is not a problem; the real signs that we must animate are by definition possible to perform. This implies that a synthetic animation system for signing does not have to solve general collision problems (which are computationally expensive), but only a few special cases, such as the elbow positioning described above.

3.3 Contacts

Besides specifying a hand position as a point in space or on the body, HamNoSys can also notate contacts between parts of both hands. The BSL two-handed spelling signs are an example of these. The inverse kinematic problem of calculating the arm angles necessary to bring the two hand-parts into contact can be reduced to two one-arm problems, by determining for each arm separately the joint angles required to bring the specified part of the hand to the location in space at which the contact is required to happen.

This is an area in which motion capture has difficulty, due to the accuracy with which some contacts must be made. A contact of two fingertips, for example, may appear on playback of the raw data to pass the fingers through each other, or to miss the contact altogether. This is due to basic limitations on the accuracy of capture equipment. When using motion capture data we deal with this problem by editing the calibration data for the equipment in order to generate the correct motion.

3.4 Handshapes

HamNoSys distinguishes around 200 handshapes. At present, we are implementing these simply by specifying the angles of all the joints of the hand for each handshape, a tedious but routine task. There is a certain amount of structure to the class of handshapes which reduces the effort: for example, a given bend of the index finger may occur in many different handshapes. For the best quality of hand shape, the joint angles should be generated algorithmically from a knowledge of the geometry of the particular avatar's hands.

4 Motion synthesis

We have so far discussed how to synthesise a static gesture. Many signs include motion as a semantic component, and even when a sign does not, the signer must still move to that sign from the preceding sign, and then move to the next.

If we calculate the joint angles required for each static gesture, and then linearly interpolate over time, the effect is robotic and unnatural. Artificial trajectories can be synthesised (e.g. sine curve, polynomial, Bezier, etc.), but we take a more biologically based approach and model each joint as a control system.

For the particular application of signing, the modelling problem is somewhat easier than for general body motion, in that accurate physics-based modelling is largely unnecessary. Physics plays a major role in motions of the lower body, which are primarily concerned with balancing and locomotion against gravity. This is also true of those motions of the upper body which involve exertions such as grasping or pushing. Signing only requires movement of upper body parts in space, without interaction with external objects or forces. The effect of gravity in shaping the motion is negligible, as the muscles automatically compensate.

We therefore adopt a simplified model for each joint. The distal side of the joint is represented as a virtual mass or moment of inertia. The muscles are

assumed to exert a force or torque that depends on the difference between the current joint angle and the joint angle required to perform the current sign, the computation being described in detail below.

Simplifications such as these are essential if the avatar is to be animated in real time.

4.1 A brief introduction to control systems

This brief summary of control theory is based on [11]. We do not require any mathematics.

A *control system* is any arrangement designed to maintain some variable at or near a desired value, independently of other forces that may be acting on it.

In general, a control system consists of the following parts:

1. The *controlled variable*, the property which the controller is intended to control.
2. A *perception* of the current value of the controlled variable.
3. A *reference signal* specifying the desired value of the controlled variable.
4. An *error signal*, being the difference between the reference and the perception.
5. The *output function*, which computes from the error signal (and possibly its derivatives or integrals) the *output signal*, which has some physical effect.

The effect of the output signal in a functioning control system is to bring the value of the controlled variable closer to the reference value. The designer of a real (i.e. non-virtual) control system must choose an output function which will have this effect. For the present application, our task is to choose an output function such that, when the reference angle for a joint is set to a new value, the joint angle changes to reach that value in a realistic manner.

4.2 Hinge joints

Our application of this avatar animation places a controller in each joint. For a hinge joint such as the elbow, the controlled variable is the angle of the joint. The reference value is the angle which is required in order to produce some gesture. The perception is the current angle. The output is a virtual force acting on the distal side of the joint. The latter is modelled as a mass, whose acceleration is proportional to the force. We also assume a certain amount of damping, that is, a force on the mass proportional to and in the opposite direction to its velocity.

The mass, the force, and the damping are fictitious, and not intended as an accurate physical model; their values are tuned to provide a realistic-looking response to changes in the reference value.

Figure 2 illustrates the response of this system to a series of sudden changes in the reference value. For a sequence of static gestures, the reference value will change in this manner, taking on a new value at the time when the next gesture is to be performed.

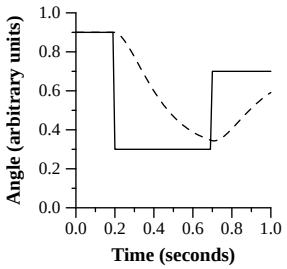

Fig. 2. Solid line: reference value. Dashed line: current value.

4.3 Higher degree joints

A turret or universal joint has two degrees of freedom. An example is the joint at the base of each finger, which can move the finger up and down, or left and right, but cannot rotate the finger about its own axis. This can be modelled as a pair of hinge joints at right angles, and the method of the preceding section applied to each hinge. This will not be accurate if the rotation of either hinge approaches a right angle, but the universal joints in the upper body all have sufficiently limited mobility that this is not a problem.

A ball and socket joint such as the shoulder has three degrees of freedom. In principle, it can be modelled by three hinge joints in series. However, there is no obvious way to choose axes for the hinges that corresponds to the actual articulation of the muscles. Mathematically, there are singularities in the representation, which correspond to the physical phenomenon of "gimbal lock", which does not occur in a real shoulder joint. Aesthetically, animation of the three hinge angles tends to give wild and unnatural movements of the arm.

Instead, we reduce it to a single one-dimensional system. If the current rotation of the shoulder joint is q, and the required rotation is q', we determine a one-dimensional trajectory between these two points in the space of three-dimensional rotations. The trajectory is calibrated by real numbers from 0 to 1, and this one-dimensional calibration used as the controlled variable. When the reference value is next changed, a new trajectory is computed and calibrated.

4.4 Moving signs

Some moving signs are represented in HamNoSys as a succession of static postures. They may also be described as motion in a particular direction by a particular amount, with or without specifying a target location. These can be animated by considering them as successive static postures, in the latter case calculating a target location from the direction and size of the movement.

HamNoSys can also specify the tempo of the motion or its path, when the path is not the default straight line motion. If the tempo is unmarked, it is performed in the "default" manner, which is what we have attempted to capture by our control model of motion. It may also be fast or slow (relative to the general

tempo of the signer), or modulated in various ways such as "sudden stop" or "tense" (as if performed with great effort). These concepts are easily understood from example by people learning to sign. Expressing them in terms of synthesised trajectories is a subject of our current work.

4.5 Sequences of signs

Given a sequence of signs described in HamNoSys or SiGML, and the times at which they are to be performed, we can generate a continuous signing animation by determining the joint angles required by each sign, and setting the reference values for the joint controllers accordingly at the corresponding times (or slightly in advance of those times to account for the fixed lag introduced by the controllers). The blending of motion from each sign to the next is performed automatically, without requiring the avatar to go to the neutral position between signs.

4.6 Ambient motion

Our current avatar has the ability to blend signing animation data with "ambient" motion — small, random movements, mainly of the torso, head, and eyes — in order to make it appear more natural. This can be used even when the animation data come from motion capture. It is particularly important for synthetic animation, since we only synthesise movements of those joints which play a part in creating the sign. If the rest of the body does not move at all — and there are few signs which require any torso motion as part of their definition — the result will look unnaturally stiff. Motion capture data can be blended with synthesized data; a possible future approach is to generate these small random movements algorithmically.

5 Target platform

These ideas were initially prototyped in VRML 97, the Virtual Reality Modelling Language [5], with a ball-and-stick avatar constructed from the H-anim specification for virtual humanoids [13] (Figure 3(a)). The control algorithms were embedded into the VRML model. Animations are thus computed at run time, not precalculated. As described above, HamNoSys handshapes were converted to joint angles by hand, and hand positions in the signing space in front of the avatar converted to arm joint angles by inverse kinematic calculations.

A ball-and-stick avatar is not suitable for production-quality signing, but it gives the animator a much clearer view of the motion. As the avatar is H-anim compliant, we were able to cut and paste several other H-anim avatars available on the Web, recompute the inverse kinematics for the new avatar's dimensions, and generate more realistic-looking animations, at the cost of reduced frame rate (Figure 3(b)).

This was useful as a prototyping exercise. However, current VRML viewers and available hardware are not fast enough to provide the frame rate of 15 to 25 frames/second required for readable signing. (25 fps is the frame rate of European broadcast television; the lower bound of 15 fps was communicated to us by Thomas Hanke.)

The avatar currently in use by ViSiCAST is called Visia, and was developed by Televirtual Ltd., one of the ViSiCAST partners. It is not H-anim compliant, but adopts a similar internal structure of a hierarchical skeleton of bones. It can be driven by a stream of information about bone rotations, positions, and lengths (although it is primarily the rotations that change from one frame to another). The avatar automatically adjusts its seamless skin to fit the bones. Visia is illustrated in Figure 3(c).

(a) Ball-and-stick avatar (b) Bob (c) Visia

Fig. 3. H-anim avatars

There is a fairly simple correspondence between H-anim joints in the torso, arms, and hands, and Visia's bones, which allows us to generate motion data for both targets with little change to the code. Visia includes many more face bones than H-anim, but synthetic facial animation is the subject of future work. A third possible target is BAP data (Body Animation Parameters), a part of the MPEG-4 standard concerned with animation of humanoid figures [6], and closely connected with H-anim.

6 Conclusions

We have described above the design and initial implementation of a method of automatic synthesis of manual signing gestures from their transcriptions in the HamNoSys/SiGML notations. We are confident that the approach described here will produce results that compare favourably with existing alternatives. Perhaps the most interesting comparison will be with the system based on motion capture, as already used in ViSiCAST. As our synthetic signing can target the same avatar model as the motion capture system, this provides the opportunity to undertake two kinds of comparison. Firstly, we will be able to do a meaningful

comparison of user reaction to our synthetic signing with reaction to signing based on motion capture. In addition, as our synthesis process drives the avatar via a stream of data whose form is identical to that produced from motion-capture, we are also in a position to perform quantitative comparisons between the two methods. In particular, if the evidence warrants it, we could consider a hybrid approach, combining synthetic generation with elements of motion-captured data.

7 Acknowledgements

We are grateful for funding from the EU, under the Framework V IST Programme (Grant IST-1999-10500). We are also grateful to colleagues, both here at UEA, and at partner institutions, for their support.

The "Bob" avatar of Figure 3(b) is available at *http://ligwww.epfl.ch/~babski /StandardBody/*, and is due to Christian Babski and Daniel Thalmann at the Computer Graphics Lab at the Swiss Federal Institute of Technology. Body design by Mireille Clavien.

References

1. D. Connolly. *Extensible Markup Language (XML)*. World Wide Web Consortium, 2000.
2. R. Elliott, J.R.W. Glauert, J.R. Kennaway, and I. Marshall. The development of language processing support for the ViSiCAST project. In *ASSETS 2000 - Proc. 4th International ACM Conference on Assistive Technologies, November 2000, Arlington, Virginia*, pages 101–108, 2000.
3. Andrew Glassner. Introduction to animation. In *SIGGRAPH '2000 Course Notes*. Assoc. Comp. Mach., 2000.
4. Jessica Hodgins and Zoran Popović. Animating humans by combining simulation and motion capture. In *SIGGRAPH '2000 Course Notes*. Assoc. Comp. Mach., 2000.
5. The VRML Consortium Incorporated. *The Virtual Reality Modeling Language: International Standard ISO/IEC 14772-1:1997*. 1997. `http://www.web3d.org/Specifications/VRML97/`.
6. R. Koenen. *Overview of the MPEG-4 Standard*. ISO/IEC JTC1/SC29/WG11 N2725, 1999. `http://www.cselt.it/mpeg/standards/mpeg-4/mpeg-4.htm`.
7. M. Lincoln, S.J. Cox, and M. Nakisa. The development and evaluation of a speech to sign translation system to assist transactions. In *Int. Journal of Human-computer Studies*, 2001. In Preparation.
8. I. Marshall, F. Pezeshkpour, J.A. Bangham, M. Wells, and R. Hughes. On the real time elision of text. In *RIFRA 98 - Proc. Int. Workshop on Extraction, Filtering and Automatic Summarization, Tunisia*. CNRS, November 1998.
9. F. Pezeshkpour, I. Marshall, R. Elliott, and J. A. Bangham. Development of a legible deaf-signing virtual human. In *Proc. IEEE Conf. Multi-Media, Florence*, volume 1, pages pp333–338, 1999.
10. Zoran Popović and Andrew Witkin. Physically based motion transformation. In *Proc. SIGGRAPH '99*, pages 11–20. Assoc. Comp. Mach., 1999.

11. W. T. Powers. *Behavior: The Control of Perception*. Aldine de Gruyter, 1973.
12. S. Prillwitz, R. Leven, H. Zienert, T. Hanke, J. Henning, et al. *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages — An Introductory Guide*. International Studies on Sign Language and the Communication of the Deaf, Volume 5. University of Hamburg, 1989. Version 3.0 is documented on the Web at `http://www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html`.
13. B. Roehl. *Specification for a Standard VRML Humanoid*. H-ANIM WG, U.Waterloo, Canada, 1998. `http://ece.uwaterloo.ca/~hh-anim/spec.html`.
14. M. Wells, F. Pezeshkpour, I. Marshall, M. Tutt, and J. A. Bangham. Simon: an innovative approach to signing on television. In *Proc. Int. Broadcasting Convention*, 1999.
15. Andrew Witkin and Zoran Popović. Motion warping. In *Proc. SIGGRAPH '95*, 1995.