Combining constraint-based models for Sign Language synthesis

Michael Filhol, Maxime Delorme, Annelies Braffort
LIMSI/CNRS
B.P. 133, 91 403 Orsay Cedex
E-mail: michael.filhol@limsi.fr, maxime.delorme@limsi.fr, annelies.braffort@limsi.fr

Abstract

The framework is that of Sign Language synthesis by virtual signers. In this paper, we present a sign generation system using a variety of input layers, separated on two sides: an anatomical side and a linguistic side. In a first part we suggest a way of implementing the flexibility required by Sign Languages into the system by using combinations of necessary and suficient constraints. The anatomical side of the input specifies all morphological and articulatory constraints that model the behaviour of a human skeleton, while the linguistic input specifies language constraints (lexical, grammatical, iconic...) that must be applied to the signer's body to utter the correct sign sequence. A second part explains how to combine all these parts of the input in a conjunction of constraints for each time frame of the animation. A point is made that conflicting constraints may be given and need be prioritised in order still to decide on acceptable solutions. A first idea of a global priority order is given to illustrate this issue.

1. Introduction & Context

Sign Languages (SLs) are the most natural way for the Deaf to communicate. Deaf people not all being comfortable with reading text, and for them to access everyday's information, we choose to combine audio information systems like station announcements with SL displays on screens. Those displays could play videos of people signing complete utterances but the nature of the information (generally flexible gap sentences) prevents us from doing so. A more flexible way of displaying SL on a screen is the use of a 3d signing humanoid called virtual signer (VS). A VS can be animated by hand, requiring professional and talented graphists, or by automatic generation, which requires all sorts of models.

Since SLs are natural languages, they have their own syntax and lexicon that need to be modelled. For the signed output to be natural and understandable by deaf people, we also need realistic models for the VS: skeleton models, animation models and skinning models.

This paper introduces a system combining several input models for the generation of signs. Section 2 addresses the models used, advocating the use of constraint-based models to synthesize signs and animate the VS. Section 3 deals with the construction of the final animation, by explaining how all parts of the total input are combined.

2. Using constraints as input for sign generation

The goal is to animate the VS with linguistically structured gesture. To carry out the task, it is therefore natural to consider at least a linguistic and an anatomical influence on the body. In this section we give an overview of the approach used for linguistic modelling in the system, then we discuss the anatomical model.

2.1 Linguistic Constraints

The linguistic side of the system generates the input coming from language-ruled principles such as lexical sign specification, grammatical structure or prosody. We presently only have a model for lexical description, called Zebedee, the grammatical layers remaining work in progress.

As we stated above, naturalness of the output animations is also a goal for the task, and the tremendous flexibility of Sign Language makes it very challenging in that respect. So far, systems generating SL from formal input (Hanke, 2002) have used phonetic descriptions like HamNoSys (Prillwitz, 1989) that specify body (in fact here, mainly hand) activity for each lexical unit (sign). Our recent work (Filhol, 2006) explains that due to the parametric structure of the approach, flexible values become rigid. In other words, in a signed sentence, every described sign results in one and only signed form, thus the flexibility of signs is not accounted for.

To provide as much flexibility as possible, our work at LIMSI has been focusing on the design of models based on sets of constraints that avoid both under- and over-specification of what needs to be uttered (Filhol, 2009).

The basic Zebedee structure of a sign is a sequence of timing units (see 'TU's on fig. 1) aligned on a timeline, where each unit specifies **everything** that is required in the period of time it covers—like a certain direction along which to align a bone or a point where to place a body site—and **only that**. In other words, a minimal conjunction of lexically intended articulatory constraints is given for each timing unit, thereby building a set of (lexically) necessary and sufficient constraints (NSCs). Then, at any moment when signing, anything left unconstrained can virtually be performed in any possible way.

The point of avoiding over-specification is to leave things open for additional constraints to be added if needed, for reasons like:

- *iconicity*: 'citation form' of lexical units are often modified according to their iconic features to fit a given context (Zebedee handles that well);
- *role shifts*: when impersonating a character with a certain body posture while uttering a sign, all unconstrained articulators can be used for the shift, leaving the lexically constrained ones for the sign;
- *grammatical reasons*: if not required otherwise by the lexicon, grammar may require that the body lean forward (e.g. a form of future in LSF), raise the eyebrows (e.g. neutral yes/no question), and so forth;

- *anatomical reasons*, which are discussed in the next section;
- etc.

Similarly, all these influences on the body are specified with as many and as few constraints as possible. They will then be combined, together with those coming from the morphology of the body.

2.2 Anatomical Constraints

Linguistic constraints are not sufficient to build a correct sign. Since one of the priorities of the generation is the realism of the final animation, we need to add a little more information. The generation of signs can be summed up as the construction of N frames in which the skeleton of the VS must be set in a particular posture. The overall generation can thus be seen as the generation of a succession of postures. For each posture, the linguistic model gives us information on how some parts of the body should be placed and oriented in space. Finding a correct posture from this information is called inverse kinematics (IK). IK problems are often under-specified problems leading to many solutions for a given input. For instance placing the wrist at a specific location in space raises an infinity of solutions (rotation of the elbow). Considering a set of possible solutions to one problem, the only element that will allow us to prefer a solution to another is the naturalness of the pose. We then add information about how realistic a posture is by informing the resolution system about the nature of the skeleton. These anatomical constraints are of three kinds:

- *joint limits* give the range of motion of each degree of freedom of the skeleton, avoiding impossible angles for the body;
- *angle probability* tells how often a specific angle of a degree of freedom is found. This measure is built from general purpose motion capture databases (Carnegie Mellon University) and a statistical analysis (Delorme, 2010);
- *biomechanical data* enhances the general quality of the posture for specific joints.This data is applied on small portions of the skeleton like the hands (Neff, 2006). Since biomechanical simulations are usually time consuming we prefer the use of pre-computed tables instead of running a real-time model.

All of these constraints apply to the skeleton and will not be subject to variation throughout the whole synthesis. Thus, in order to generate the animation, we consider on one side constraints coming from a linguistic point of view, that define what is mandatory for the sign or sentence. On the other side, we look at constraints that apply to the body and stay constant through time. We are now going to see how these two kinds of constraints interact in the generation system for sign synthesis.

3. Combination of constraints

Using all these linguistic and anatomical constraints allows us to reduce the number of possible solutions, and eventually choose one as the best posture for a given problem (i.e. one frame). Figure 1 illustrates the layers of constraints generated by the different models mentioned in section 2, and what we mean by conjunction of constraints for each time frame.

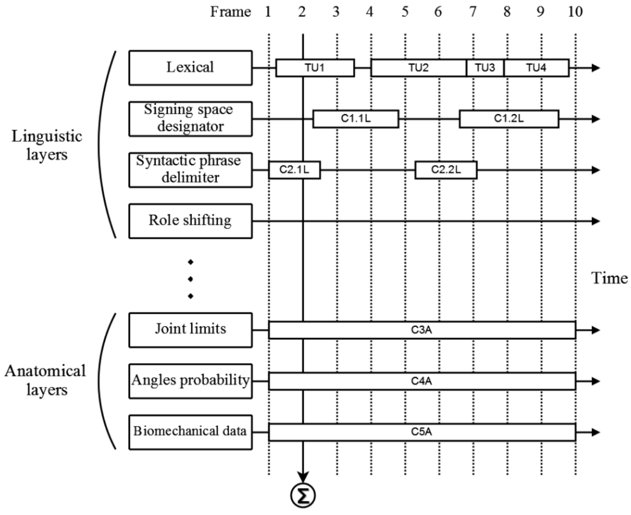

Figure 1: Resolution of multiple constraints through time

On the vertical axis we enumerate the layers of language (upper part of the list) and of anatomy (lower part) that may raise constraints on body articulations when signing. For instance, the purpose of the layer named "signing space designation" is to act on eye gaze and head (body articulators) as required in LSF to activate relevant parts of space or locate a new object by directing those articulators to the relevant points in space. The "syntactic phrase delimiter" will act on eyebrows and shoulders to mark topics in LSF, eyebrows for interrogatives, probably do some head shaking to emphasise negative clauses, etc. In the case of a dialog, "role shifting" will turn the body into the right direction to account for the alternating speakers. This layer will also use arms or hunch the back when impersonating characters with such distinctive markers.

We left the list of layers open as we imagine any number of them can be added to include more features, either additional language-specific rules, discourse prosody or indeed signing style, etc.

Theoretically, while the addition of constraints simply specifies the IK problem more (moving it away from under-specification), it also increases the risk for the problem to become over-specified (no solution).

A timeline is attached to each of these layers. In the diagram, time flows from left to right. When a layer generates a set of constraints over a period, they are represented by a white box on the timeline. As we said earlier, anatomical constraints remain constant in time, which is why the bottom lines have a box covering the whole animation without a change. At this point, it is clear that the set of constraints applying to the body at any moment in time is the conjunction of all constraints present on all layers at that moment.

Time is then broken in a sequence of frames to generate the output video. These time frames are shown across the drawing and numbered at the top, representing where to take snapshots of the timelines, each snapshot raising the set of constraints to combine hence a problem to solve for the time frame. On our example, frame no. 2 involves lexical constraints (from block TU1) and syntactic constraints (from C2.1L, say to mark the lexical sign as a sentence topic), as well as all anatomical constraints (C4A, C5A and C6A). It bears no space designation constraint for instance, the first frame where these occur being frame no. 3, from block C1.1L.

While all constraints are given equal consideration, they may be processed in different stages of the synthesis. Constraints can be set to ask for contradictive or conflicting orders if two of them are located on the same parts of the skeleton. A good example of such conflict would be in French Sign Language (LSF) to sign "I know" while role shifting in a wolf character as illustrated in figure 2. To look more frightening, the signer frowns, hunches his back, raises his elbows and puts his hands (paws) forward. But to sign "I know", the signer needs to bring his strong hand to his forehead. So the system is given two orders regarding the right arm. There is no definite way of solving such conflicts since the priorities are sign-dependant. We chose to: first, give arbitrary priorities to the constraints, even if we know that this is not a really satisfactory solution; second, segment the skeleton into independent parts that will, to some extent, behave separately.

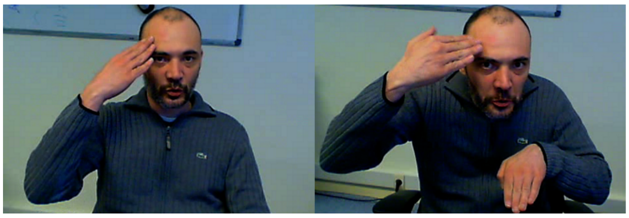

Figure 2: Left, "I know" in LSF; Right: the same sign while role-shifting as a wolf.

Here is an example of a simple priority scheme for constraints, based on the intuition of "what will work more often". This part of the work will of course need more investigation.

1. Joint limits are the absolute priority. We cannot have the VS make impossible angles.
2. Lexical constraints follow. They define as stated before what is absolutely necessary in the sign.
3. Grammatical layers add important information on the signs and must then be considered. Angle probabilities allow the system to choose in the resulting a set of solutions.
4. Finally, biomechanical data improves the configuration of unconstrained effectors (e.g. fingers) regardless of what has been previously computed.

The segmentation of the skeleton in five parts (see fig. 2) allows us to locate precisely which bone of the skeleton should be considered for a single problem. Thus a problem considering the right elbow will only involve the section "right arm", leaving the other parts free to be affected by different constraints. This might not be sufficient. For instance, a sign like [TREE] in French Sign Language needs the signer to place his weak hand on a specific location in space. Thus, considering only the hand from the wrist will fail. When no satisfactory solution is found to a problem, the system tries again the resolution with a longer kinematic chain (i.e. a sequence of bone of the skeleton to move). In the precise case of [TREE], the system will consider the hand and the arm at the same time. If it still is not sufficient then the system will consider the complete kinematic chain including the hand, the arm and the spine (for instance signs needing to place the hand far from the body will lean the body forward to reach out further).

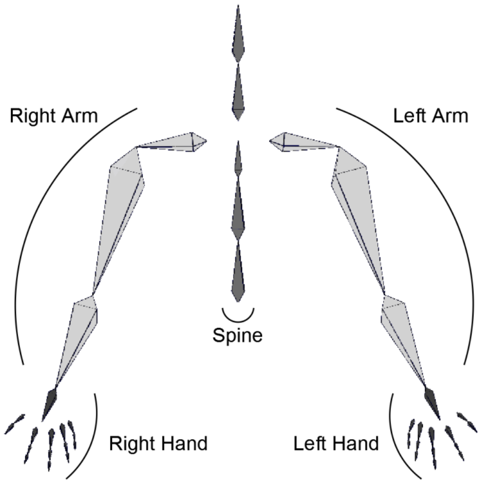

Figure 3 : Segmentation of the skeleton for progressive IK

The core of the resolution is the IK module. IK is a well-known problem in robotics (Lee, 1993) and animation (Komura, 2005). It consists in finding rotation angles for a kinematic chain to place an effector (e.g. the wrist, a finger, the elbow) at a specific location in space, or to orient it in a specific direction. The method we choose to solve IK problems is based on sequential Monte-Carlo simulations (Courty, 2008). This method is preferred to more common ones (Wang, 1991; Maciejewski, 1990) because of its very narrow connexions with probability distribution functions allowing us easily to include the anatomical constraints. The adaptation to our case works as follow:

1. We generate a certain number of random configurations for our skeleton. The range of the random angles is set to remain within the joint limits. Moreover, this generation follows the distribution functions of the angle probabilities to give more realistic results.
2. Every single solution is given a score depending on the quality of the result: in case of a placement the score depends of the distance between the effector and the target; in case of an orientation the score depends of the angle between the current orientation and the target orientation.
3. Each solution moves randomly around its current position trying to enhance its quality.
4. Biomechanical calibrations are made on the unprocessed parts of the skeleton to improve the overall posture.

The process iterates a limited number of times and stops if a good solution (given a threshold) is found. From this process we extract the ten best results and assign scores to them, based on the angle probabily tables. The more a configuration is found in the motion capture database, the higher its score. Finally we decide the most realistic solution is the one with the highest score and keep it as final result for the generation. This overall method is applied for each frame of the animation to generate.

4. Conclusion

We have presented a sign generation system based entirely on conjunction of constraints, coming from different layers of (for now, at least) language or anatomy. These constraints all apply to the skeleton of the VS but are synchronised differently in time according to the layer they belong to. The conjunction of all these constraints minimally specifies a posture for the skeleton at a specific time. As this can lead to conflicts, the constraints must be given relative priorities and a first tentative scheme was proposed. It should however be redefined from a precise analysis of which layers dominate the others, and indeed of whether they do constantly or in what way the scheme varies over time if not.

Such a system avoids too strong a separation between roles of articulators, e.g. dedicating the hands to the lexicon; the eyes to space activation and reference, and the torso to, say, role shifts. We separate the origins of the constraints in what we have called ‘layers’ of the system rather than what the constraints apply to. Now all layers may each act on all articulators.

Further work is needed to implement the system, as we currently have only anatomical and lexical constraints combined, but we hope this design brings to the field of sign generation more of, and an original approach to, the flexibility required by SLs.

5. Acknowledgements

The reseach leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n°231135.

Some of the data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

6. References

Carnegie Mellon University, Graphics Lab Motion Capture Database: http://mocap.cs.cmu.edu/

Courty, N., Arnaud, E. (2008). Sequential Monte Carlo Inverse Kinematics v3. *INRIA Internal Report, RR-6426,* February 2008.

Delorme, M. (2010). Sign Language Synthesis: Skeleton modelling for more realistic gestures. *SIGACCESS Newsletter,* February 2010.

Filhol, M., Braffort, A. (2006). A Sequential approach to lexical sign description. In *LREC 2006 - Workshop on Sign Languages*, Genova, Italy.

Filhol, M. (2008). Modèle descriptif des signes pour un traitement automatique des langues des signes, *PhD thesis,* Université Paris-11 (Paris sud), Orsay.

Filhol, M. (2009). Zebedee: a lexical description model for Sign Language synthesis. *LIMSI internal report* 2009-08*,* Orsay.

Hanke, T. et al. (2002). VISICAST deliverable D5-1: interface definitions. *VISICAST project report*.

Komura, T., Ho, E.S.L, Lau, R.W.H. (2005). Animating reactive motion using momentum-based inverse kinematics. *Computed animation and virtual worlds*, vol. 16, pp. 213--223.

Lee, S., Kim, S. (1993). Efficient inverse kinematics for serial connections of serial and parallel manipulators. In *Proceedings - RSJ/ICIRS, IEEE, Yokohama, pp. 1635-1641.*

Maciejewski, A.A. (1990). Dealing with ill-conditioned equations of motion for articulated figures. *IEEE Computer Graphics Applications,* vol. 10, pp 233-242.

Neff, M., Seidel, H-P (2006). Modeling Relaxed Hand Shape for Character Animation. *Articulated Motion and Deformable Objects (AMDO)*, vol. 4069 of LNCS, pp. 262--70.

Prillwitz, S. et al (1989). HamNoSys version 2.0 - Hamburg Notation System for Sign Languages, an introductory guide. *Internation studies on Sign Language and communication of the Deaf*, vol. 5. Signum Press, Hamburg.

Wang, L.T., Chen, C.C. (1991). A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. *IEEE Trans. Robotics Automation*, vol. 7, pp 489-499.