Central European Multimedia and Virtual Reality Conference (2006)
C. Sik Lányi, B.Oláh (Editors)

On the timing of gestures of a Virtual Trainer

Zs. Ruttkay and H. van Welbergen

HMI, Dept. of Compter Science, University Twente

Abstract

A Reactive Virtual Trainer is a virtual human capable to demonstrate physical exercises, to monitor the performance of the user and react in speech, both addressing the correctness of the motion, the overall performance and the motivation of the user. In this paper we focus on the timing of the simple exercises motions (called gestures). We discuss the subtle requirements on synchronizing motion to speech and other acoustic modalities, and the proper timing of the stages of the gesture. We outline our implemented demo, to visualize and to experiment with the different planning and timing strategies. Finally, we talk about further work concerning motion models and real-time adaptive planning for our RVT application.

Categories and Subject Descriptors (according to ACM CCS) I.3.3 [Computer Graphics]: Animation, embodied conversational agent, VR

1. Introduction

A Reactive Virtual Trainer (RVT) is an Embodied Conversational Agent (ECA) [1] in the role of a human trainer, capable of monitoring and coaching a real person performing physical exercises. The challenges of the research we have just started [11] are in making the RVT believable, professionally and socially competent and reactive, which are prerequisites for its effect in a real application. The RVT, similar to a real expert, observes the performance of the human, and reacts as needed: by pinpointing inaccuracies of the exercise as performed by the user and by demonstrating the correct motion, by noticing if the user is getting exhausted or loosing motivation and reacting by shortening the exercises, or by introducing extra relaxing times. These professional reactions should be interwoven with emphatic acknowledgements, encouragements and eventual small talk, which form a substantial part of the arsenal of real physiotherapists keeping their patients motivated. Our envisioned RVT in its reactivity, multimodal capabilities and adaptability outperforms the related works [2, 4, 5, 6, 12, 13], for more detailed comparison see [11].

In this paper we focus on a specific aspect of the multiple services of a RVT, namely the presentation of physical exercises. The units of repetitive exercises can be defined in a compositional way [10]. Here we address the timing, synchronisation and planning of rhythmic exercises motions. We will refer to ‘units’ like a clap, or a single arm pull as a *gesture*. However, they are different of, and less investigated, as the usual gestures accompanying speech. On the other hand, the subtle timing and proper performance of the exercises is of primarily importance, as opposed to the general principles to be followed and variations allowed in case of communicative gestures.

In Section 2 we discuss the timing and synchronisation requirements of rhythmic exercises. In Section 3 we outline our multi-modal planning system and illustrate the approach we take with an implemented demo environment for clapping. Finally, in Section 4 we explain further work building on the present model and experimental environment.

2. Generating rhythmic exercises

A rhythmic exercise consists of a single gesture performed several times, in a repetitive way. In case of training, the tempo is often given by music, in an implicit way, or by tapping or clapping, or counting, in an explicit way. Counting is also used to keep track of the progress, and to indicate the ‘peak’ of the motion. The major modelling and computational tasks related to these phenomena are discussed on by one.

2.1. Multi-modal Synchronization

If an ECA is to converse in usual way, the speech-accompanying gestures are synchronised to the speech. In the case of an RVT, there may be further acoustic modalities to be taken into account, namely music, tapping or clapping (of the RVT, or of the user), besides speech. On

each modality, *unit actions* of that modality (a word, a piece of recorded sound, a hand gesture) can be executed. A unit action has a set of predefined moments of synchronization, called *alignment points*. The start and end of any unit action are alignment points, but there may be additional ones such as the start of the emphasized syllable of a word, or the start of phases of hand gestures (see 2.2). Synchronization is defined solely by prescribing the alignment of some of the alignment points between units of different modalities. Prescribed alignments pose constraints on durations of (phases of) unit actions. In order to meet these constraints when generating an actual plan of execution of actions, shifting the start time of a unit and/or scaling up/down the duration of phases may be needed. It is a characteristic of a unit action which phases and to what extent may be scaled. E.g., hand gestures may be slowed down by holding the hand still between movement phases, and/or by slowing down movements of different phases.

2.2. Timing of a gesture

For exercises unit gestures, we adopt the terminology used for the stages of communicative gestures:

- ? preparation, indicating the motion to take the 'start' posture for the gesture;
- ? stroke, indicating the major, energetic motion of the gesture;
- ? retraction (optional), taking the hand back to a resting or idle position;
- ? pre-hold, and post-hold intervals indicating that the hand or body remains for some time in the posture of preparation or at the end of the stroke.

In case of repetitive gestures, each stroke is followed by a motion which is the preparation of the next stroke, hence these 2 motions with optional holds between them, form a unit (see Figure 1).

For timing a gesture of a given duration (according to prescribed tempo), the time spent in still, hold positions and the distribution of the remaining time between the two stages of motion need to be decided. The proportions may not be the same for different tempi. Moreover, there are limitations (may be different for different users) on the upper-lower speed of a gesture, as well as on the duration of the stages. These, may be changing, characteristics need to be taken into account when planning the gestures automatically.

Another subtle aspect is the relationship between the length and shape of the path the moving hand, and the speed profile of the motion. The former will be taken care of by the definition of the gesture, while the latter by the planner. It may be necessary to decrease the 'amount' of motion, that is, the amplitude of the gesture, in order to be able to perform it in a shorter time.

Yet a subtle case is the ending of an exercise, when the motion often slows down, and a resting position is reached.

2.3. Real-time reactive planning

The RVT keeps monitoring the user, and reacts according to his/her performance. E.g. if the user does not keep up with the tempo, the RVT may decide to emphasize the tempo by counting, or to adjust his own tempo to the users' who seems to be very exhausted, or may decide to abandon the exercise altogether. These intelligent decisions result in new schedules of exercises possibly accompanied by speech, and new tempi. Hence the motion of the RVT needs to be re-planned in a reactive way, and in real-time.

Also the motion between different exercises – preparing the next one, or taking a resting position, may be with some idle motion – should be generated automatically, by the planner.

3. Clapping demo

We have implemented a demo system to experiment with the timing and planning strategies of rhythmic exercises. We used clapping as demo exercise, where the clap sound should be generated at the right time, and the clapping may be accompanied by counting speech.

We used our parameterized animation developed for the virtual dancer to create a clapping gesture [9]. Clapping can be performed at different amplitudes. At higher amplitudes, the hands start out further from each other and more of the body is involved in the clapping action: the head nods and the upper-body sways along with the clap.

Speech is generated using the Loquendo speech generator [8]. It allows us to identify the timing of the most stressed phoneme, which we use to align the words to the gesture. The clap sound is generated using a .wav file. The entire system is written in java, using java 3D. Our experiments run so far in real time on standard pc under Windows.

3.1. Definition of exercises in script

BML is a multimodal generation language, describing synchronization between speech and animation on such a level that it can be used as input for the final process of multimodal generation [7]. We extended BML to allow explicit sequential relations between behavior. Speech markers are extended in such a way that other behavior can align to them at word start/end or phonological peak start/stroke/end). Observers are added to monitor outside behavior (for example, the beats of music) and align behavior to that. In Figure 3 you see the definition of two clap sequences.

3.2. Timing of clapping

We have developed an interactive environment where the user may specify explicitly the tempo of claps, and tell the tempo of preparation and stroke, and how to distribute the remaining time (if any) between holds before and after the clap stroke. The character will perform the clap sequence according to the specification, see Figure 2. Currently, there is no mechanism to prevent or correct inconsistent,

infeasible specifications. Partial specifications will be extended according to the 'default' strategy. E.g. if the tempo is to be slowed down, the duration of preparation and stroke will be slowed down proportionally, as default. However, the user may specify, for instance, that in the slow tempo the stroke should not be so slow, and the remaining time is assigned automatically to a hold, see Figure 2.

3.3. Synchronization

The clap gesture has five alignment points, that are linked to positions in the animation: the start of the animation, the end of the animation, the stroke, the post-stroke-hold, the post-retraction-hold and the end. The stroke and post stroke hold and the post-retraction-hold and the end are linked to the same position in the animation, respectively. To perform the animation, all these alignment points have to be mapped to real time values, time warping the animation. By mapping the post-stroke-hold alignment point to a later time value than the stroke alignment point, the clapped hands remain still for some time (see Figure 1).

We use a metronome to generate absolute synchronization points for the clap exercise. The acoustic modalities used are clapping sound and speech. Hence altogether 3 modalities are to be synchronized, using the metronome beats as reference times. The (uneven) tempo of clapping, as well as the subtle coordination of the speech and the hand motion are given as BML script (see Figure 3). From this script, the actual plan is generated (see Figure 4).

4. Further work

The current environment allows systematic experimentation. We plan to test the perception of the different timing and sync strategies. We wish to explore the boundaries of 'natural' clapping, the range of possible individual differences, and the manifestation of emotional content (enthusiastic clap versus polite clap).

Another source for restricting the parameters to a smaller region of natural gestures is the human behaviour. We shall continue our work to gather data on individuals clapping by using motion capture technology. Such data will also allow to refine the path-speed profile of claps, and gain insight into the correlation between speed and amplitude.

Our planner needs to be improved, concerning the handling of constraints. Also, concatenation of gestures, automatic introduction of rest positions and idle motions should be assured.

An open issue is how to tackle speech tempo changes. The current TTS system we work with does have a tempo parameter, however, this does not mean exact scaling of the duration of the utterance. We must find a way around, in order to be able to adjust speech to a given tempo.

At a later stage, we plan to align an exercise to the music using our beat tracker [3].

References

1. J. Cassell, J. Sullivan, S. Prevost and E. Churchill (eds.), *Embodied Conversational Agents*, MIT Press, 2000.
2. T. Bickmore, R. Picard: *Towards caring machines*, Proc. of CHI, 2004.
3. P. Bos, D. Reidsma, Z.M. Ruttkay, A. Nijholt: *Interacting with a Virtual Conductor*, Proc. of 5th International Conference on Entertainment Computing, *LNCS* 4161, Springer Verlag, pp. 25-54, 2006
4. S-P.Chao, G-Y Chiu, S-N, Yang, T-G. Lin: *Tai Chi synthesizer: a motion synthesis framework based on key-postures and motion instructions*. Computer Animation and Virtual Worlds, Vol. 15. pp. 259-268, 2004.
5. J. Davis, A. Bobick: Virtual PAT, MIT Media Lab Technical Report 436, MIT, 1998.
6. W. IJsselsteijn, Y. de Kort, J. Westerink, M. De Jager, R. Bonants: *Fun and Sports: Enhancing the Home Fitness Experience*, Proc. of ICEC 2004.
7. Kopp, S., B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson: *Towards a Common Framework for Multimodal Generation: The Behavior Markup Language*. Proc. of IVA 2006, LNAI 4133, Springer-Verlag, 2006
8. Loguendo TTS: www.loquendo.com
9. D. Reidsma, H. van Welbergen, R.W. Poppe, P. Bos. A. Nijholt: *Towards Bi-directional Dancing Interaction*, Proc. of 5th International Conference on Entertainment Computing, LNCS 4161, Springer-Verlag, *2006*, pp. 1-12, *2006*,
10. Zs. Ruttkay, Z. Huang, A. Eliëns: *Reusable Gestures for Interactive Web Agents*, In: R. Aylett, D. Ballin, T. Rist (Eds.), Intelligent Virtual Agents, Proc. of IVA-2003, LNAI 2792, Springer-Verlag, pp. 80-87.
11. Zs. Ruttkay, J. Zwiers, H. van Welbergen, D. Reidsma: *Towards a Reactive Virtual Trainer*, Proc. of IVA 2006, LNAI 4133, Springer-Verlag, pp. 292-303. 2006
12. Sony: EyeToy: Kinetic, http://www.us.playstation.com/Content/OGS/SCUS-97478/Site/ and http://www.eyetoykinetic.com
13. J. Westerink, M. de Jager, M., Y. de Kort, W., IJsselsteijn, R., Bonants, J. Vermeulen, J. van Herk, M. Reersma: *Raising Motivation in Home Fitnessing: Effects of a Virtual Landscape and a Virtual Coach with Various Coaching Styles*. Proc. of ISSP 11th World Congress of Sport Psychology, 15 - 19 August 2005, Sydney, Australia.

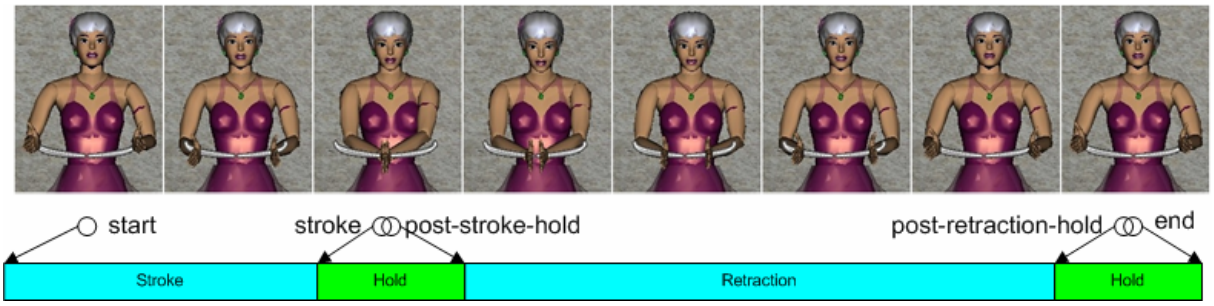

Figure 1: *Stages of a repetitive gesture, such as clap.*

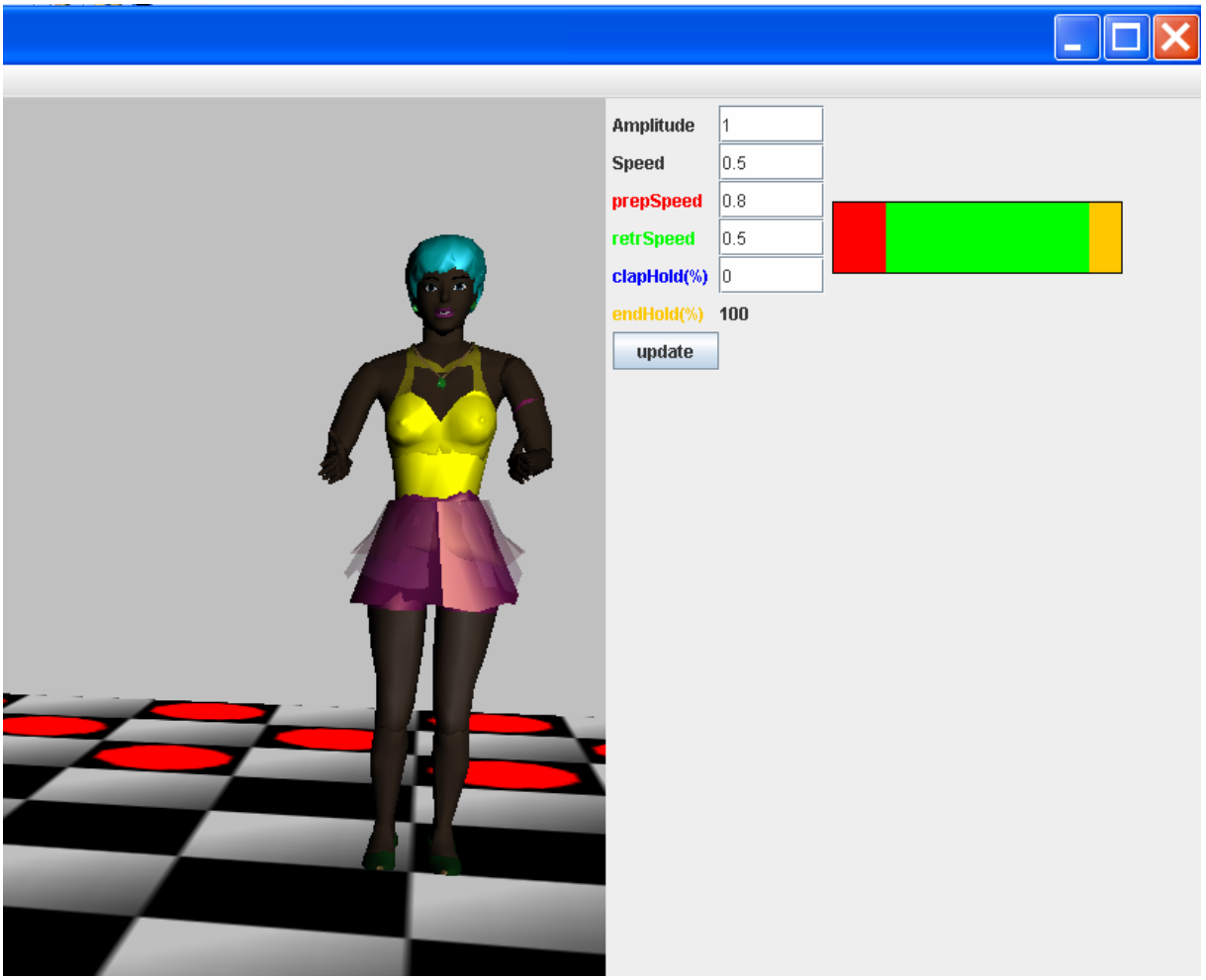

Figure 2: *Snapshot of part of the interactive environment to specify and generate clap sequences of different timings .The user-specified timing for the stroke and retraction stages is automatically extended with a hold at the end.*

```
 <bml>       <observer id="beatObserver1"/>
  <speech id="one" stress:strokePhonPeak="beatObserver1:  1"
type="application/sapi+xml"><bookmark mark="stress"/>one</speech>
  <speech id="two" stress:strokePhonPeak="beatObserver1:2"
type="application/sapi+xml"><bookmark mark="stress"/>two</speech>
  <speech id="three" stress:strokePhonPeak="beatObserver1:4"
type="application/sapi+xml"><bookmark mark="stress"/>three</speech>
  <speech id="four" stress:strokePhonPeak="beatObserver1:6"
type="application/sapi+xml"><bookmark mark="stress"/>four</speech>
  <speech id="five" stress:strokePhonPeak="beatObserver1:7"
type="application/sapi+xml"><bookmark mark="stress"/>five</speech>
  <animation stroke="one:stress:strokePhonPeak" id="clap1" name="clap_fast"/>
  <animation stroke="two:stress:strokePhonPeak" start=">clap1:end" id="clap2"
name="clap_fast"/>
  <animation start="cl ap2:end" id="interpolator1" name="interpolator"/>
  <animation stroke="three:stress:strokePhonPeak" start=">clap2:end" id="clap3"
name="clap"/>
  <animation start="clap3:end" id="interpolator2" name="interpolator_rev"/>
  <animation stroke="four:stress:stroke PhonPeak" start=">clap3:end" id="clap4"
name="clap_fast"/>
 <animation stroke="five:stress:strokePhonPeak" start=">clap4:end" id="clap5"
name="clap_fast"/>
  <sound id="clapsound1" name="clap" clap="clap1:stroke">
      <tm id="clap" time="0.02" /> </sound>
  <sound id="clapsound2" name="clap" clap="clap2:stroke">
      <tm id="clap" time="0.02" />  </sound>
  <sound id="clapsound3" name="clap" clap="clap3:stroke">
      <tm id="clap" time="0.02" />  </sound>
  <sound id="clapsound4" name="clap" clap="clap4:stroke">
      <tm id="clap" time="0.02" />  </sound>
  <sound id="clapsound5" name="clap" clap="clap5:stroke">
      <tm id="clap" time="0.02" />  </sound>
 </bml>
```

Figure 3: *The script of a clapping sequence. Claps are accompanied by speech, are performed at different times at 1,2,4,6 and 7, and of different durations.*

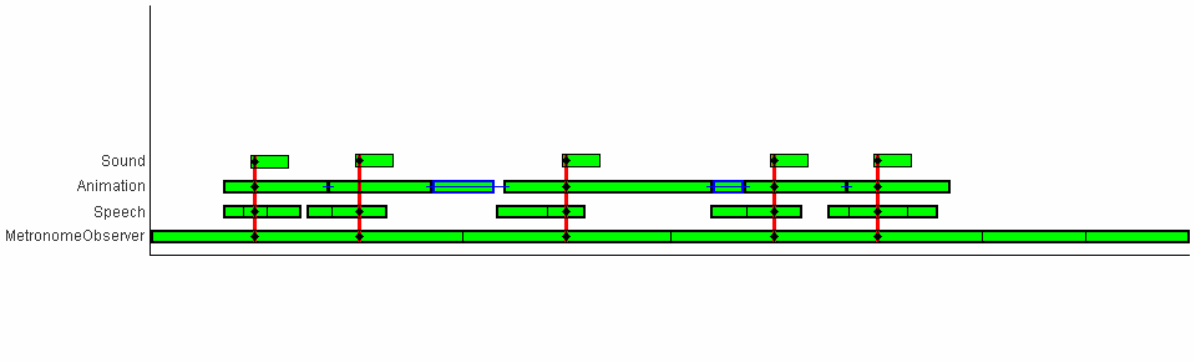

Figure 4: *The visualization of the plan generated according to the above specification, with synchronization of 4 modalities: hand motion, speech, clap sound, and a metronome defining tempo.*