Universal Access in the Information Society manuscript No.
(will be inserted by the editor)

Hybrid Paradigm for Spanish Sign Language Synthesis

Fernando López-Colino · José Colás

Received: September, 11^{th} 2009 / Accepted: December, 16^{th} 2010

Abstract This work presents a hybrid approach to sign language synthesis. This approach allows hand-tuning the phonetic description of the signs, focusing on the time aspect of the sign. Therefore, we keep the capacity of performing morphonological operations, like the notation-based approaches and improving the synthetic signing performance, like the hand-tuned animations approach.

Our approach simplifies input message description using a new high level notation and storing sign phonetic descriptions in a relational database. The relational database allows more flexible sign phonetic descriptions; it also allows describing sign timing and the synchronization between sign phonemes. The new notation, named HLSML, focuses on message description; it is a gloss-based notation. HLSML introduces several tags that allow modifying the signs in the message defining dialect and mood variations (both defined in the relational database) and message timing (transition durations and pauses). We also propose a new avatar design that simplifies the development of the synthesizer and avoids any interference in the independence of the sign language phonemes during the animation.

The obtained results show an increase of the sign recognition rate compared to other approaches. This improvement is based on the active role that the sign language experts have in the description of signs, allowed by the flexibility of the sign storage approach. The approach will simplify the description of synthesizable signed messages so the creation of multimedia signed contents will be easier.

Keywords Spanish Sign Language · Sign Language Synthesis · Graphical User Interfaces · Human Computer Interaction

Fernando López-Colino, José Colás
Human Computer Technology Laboratory
Universidad Autónoma de Madrid
Francisco Tomás y Valiente 11, 28049 Cantoblanco, Spain
Tel.: +34-91-4973613
Fax: +34-91-4972235
E-mail: [fj.lopez; jose.colas]@uam.es

1 Introduction

Spanish Government enacted a law in 2007 which binds official documents and web sites to be accessible using Spanish Sign Language (LSE). After three years, this law cannot be applied. Machine translation systems do not provide the required flexibility and correctness on the translations to be faithful. Providing signed contents by means of recorded videos is too expensive because every new message must be recorded using sign experts and cannot be reused. Automatic synthesis allows contents reutilization, but synthetic messages are not easily accepted by deaf people. These automatic synthesizers are not flexible enough to allow sign language (SL) experts to modify the results in order to improve their quality. These systems also have another drawback: the input notation. These notations are low level sign descriptions; these descriptions focus on the Phonologic Parameters (PPs), conceptually related to the speech phonemes, that constitute the sign. These notations are quite resourceful for the definition of human static gestures. Although they can describe movements, these notations do not describe sign duration nor PPs timing, both of them are essential for correct sign representation. Finally, manual message creation requires deep knowledge of SL syntax, grammar, prosody and phonology; the complexity of these notations makes the manual definition of signed messages a difficult task.

The main motivation of our work is to improve the usability of SL synthesizers and to increase the quality of synthetic signed messages so the signs are easily recognized and identified. We have mentioned that the main problems of current synthesizers are the complex input notations that do not describe the temporal aspect of the signs and their lack of flexibility. Our proposal uses a new high level input notation, HLSML, which has been developed focusing on message description. This notation allows defining the message focusing on SL syntax, as the PPs of the signs are described independently. HLSML defines several tags to allow modifications to the sequence of elements that compose the sentence, leading to different synthetic messages. It allows reusing the same message for different LSE dialects because the difference between these dialects is the sign's PPs description, not the message's syntax. The notation also includes the possibility of defining prosodic modifications, pauses between signs and the duration of the transitions. These two aspects have proven to increase the quality of the synthetic SL messages, as stated by Huenerfauth [18].

We have stated that sign phonetic descriptions must include a temporal description including duration, PPs timing and synchronization [29]. The best approach for storing this information is a relational database whose structure observes the independence of each PP. The flexibility of the database allows hand-tuning the sign phonetic descriptions. Therefore, keeping the flexibility and capabilities of the linguistic approach used in the notation-based works, we can improve the performance of the signing avatar, in a similar way as hand-tuned animation-based works allow. This can be defined as a hybrid approach between the hand-animated approaches [42, 43, 49] and the notation-based approaches [7, 10, 52]. The structure of a relational database also makes multiple definitions of the same concept possible, which is quite useful for dialectal and mood variations. The descriptions of the signs are stored in the relational database using a specific application. This application allows the management and description of the phonologic segments that compose a sign without using any notation.

During the development of this different approach, we realized that the avatar's design proposed in other SL synthesis projects involve increasing the complexity of the gesture synthesis algorithms. We propose several modifications to the signing avatar design in order to simplify these calculations. Finally, we also present the synthesizer's modular architecture in order to adapt the synthesizer to different devices.

This paper is divided in the following sections: Sec. 2 presents a review and a discussion of the related work, Sec. 3 is a briefly introduction to the linguistic elements of SL that must be considered in the LSE synthesis. We describe the elements of our proposal focusing on the relational database in Sec. 4, the input notation in Sec. 5, the avatar's design in Sec. 6 and the architecture of the system in Sec. 7. Finally we present the objective tests in Sec. 8 and the user evaluations in Sec. 9. Sec. 10 summarizes this work and Sec. 11 contains our future work.

2 Sign Language Synthesis Related Work

Even though this work concerns gesture synthesis for sign language representation, we will briefly review complete translation systems. The ViSiCAST [1] and eSign projects [52] represent great advances in sign language translation from voice or text [7, 27]. These projects use the HamNoSys notation [12, 34] as precursor to gesture synthesis. Both projects use the same sign language synthesis module, which will be reviewed next. Moreover, LSE machine translation is in its early development steps. There is a small number of related works, which represent the first efforts on this subject [36, 37]. However, San Segundo et al. use the eSigns's synthesis module adapted to LSE signs as there is no previous work dealing with LSE synthesis.

In order to represent SL messages, several techniques have been developed using voice strategies as reference.

A first approach to SL synthesis consists of creating a composition of small segments of video. These pre-recorded elements can be played in sequence to represent a message. Video segments can represent an isolated sign, a small phrase or a whole message. Obviously, the last option cannot be defined as synthesis, but it is used in several web pages[1]. The final message is synthesized creating a sequence of pre-recorded chunks. A first approach does not include smooth transitions between consecutive videos. In order to improve final message quality, the transitions between video units can be generated using morphing techniques [44]. This approach to SL synthesis requires image processing and a great number of pre-recorded sequences in order to act as a synthesizer, and thus significant storage capacity.

The second main approach is pure SL synthesis using virtual avatars. The avatar is a 3D generated human model animated using a bone structure. Albeit with different skeleton structures, many projects [1, 7, 9, 21, 22, 52] use a similar approach to gesture synthesis. The most widely used skeleton structure is H-Anim [25], a standard definition for human representation on VRML [23, 24]. The ViSiCAST skeleton structure is very similar to that of H-Anim, but the ViSiCAST project has designed its own structure. Many projects use VRML as their graphic API, so a VRML viewer must be installed on final user devices.

Both the ViSiCAST avatar and the H-Anim definition define the position of several anatomic references (e.g. center of the chest, facial elements, ...) defining the nearest mesh's vertex. These approaches must handle the mesh deformations during the gesture synthesis process in order to obtain the correct coordinates of the relevant anatomic references.

[1] www.signwriting.com or www.cervantesvirtual.com

Most SL synthesizers use standard notations for sign phonetic description. Notations such as HamNoSys [12, 34] and SignWriting [48] are graphic representations of the PPs (see Sec. 3) and have computer-friendly versions: SiGML [6] for the HamNoSys notation and SWML [35] for the SignWriting notation. Gesture synthesis for these projects is a direct conversion from SWML or SiGML into VRML. For this reason, representation potential is related to SiGML and SWML definition. The SiGML notation allows extremely detailed definitions of gestures, but this notation does not allow defining the timing of the sign or mood modifications:

- The definition does not include the duration of the sign. The ViSiCAST project algorithmically estimates the duration of the sign using the length of the HamNoSys definition which cannot be applied to all signs.
- The HamNoSys notation does not specifically define the synchronism between different phonemes; the definition of the Non-hand parameter synchronism is especially complex. This notation states the initial values for the PPs and the actions that modify these initial values during the performance of the sign.
- These notations do not define the acceleration between different units in the same PP. The acceleration represents signer's mood and can be used to emphasize a sign within a sentence.

Other projects, also based on HamNoSys notation for describing the signs, use a different approach. Instead of using SiGML, Fotinea et al. define a module that transforms the HamNoSys definitions into the STEP notation [15][2]. This synthesis module was used for an educational application [26]. The report presented by van Zijl describes the work in progress related to a South African Sign Language Machine Translation System [51]. This work also uses the STEP notation for their synthesis module.

There is another approach for SL synthesis using avatars. This other approach uses manually-defined animations of single signs [43, 49]. A related work created handmade animations of the coarticulation between every two letters of the fingerspelling alphabet; the resulting animations were rendered into video files and composed in order to spell words [42]. Although the quality of the animation obtained with the hand-tuned animation approach can be superior to the notation-based approaches, the lack of the phonetic descriptions prevents these systems from handling the complex morphologic variations of SLs.

2.1 Synthetic SL messages evaluation

As for SL synthesis evaluation, previous works attempted to evaluate the whole translation system by considering sign quality and interaction complexity with the avatar. We have only evaluated the sign recognition rate of the synthesized signs.

The recognition rate of ViSiCAST system is 81% for isolated signs and 61% for complete phrases [5]. The experiments allowed three views of each video. The testing group was formed by six people who were born profoundly deaf and assisted by three

[2] This notation is a general purpose animation notation that allows defining the animation of a H-Anim compliant avatar. This notation focuses on joint rotation definition, so it presents a lower level of abstraction compared to SiGML.

clerks experienced in serving deaf customers. Testing results indicate that this evaluation is a deeply subjective process. SL synthesis acceptance rate is lowered by the dialectal diversity.

Huenerfauth [18] reported the relation between signing speed and pauses with the recognition rate of the synthetic message. Although we have not tested this aspect of synthetic messages, we want to stand out the size of the testing group who performed their evaluations: twelve deaf users. The size of the testing group in both the ViSiCAST project and the Huenerfauth's evaluations shows the difficulty of finding suitable users for evaluating this kind of projects.

3 Sign language linguistic work

In this section we provide a brief review of the phonological theories of SL that are used for our synthesizer and a description of how we have adapted them to the SL synthesis.

Although signs were considered as indivisible units, studies of SL theory have evolved over the last fifty years [2, 33, 45, 46]. These studies pointed out how each sign is composed by different and independent Phonologic Parameters (PPs). The number of PPs has increased from three in early studies up to seven in recent works. We have based our linguistic approach in the seven PPs theories, but we have introduced some differences in the management of some of them. Next we provide a brief discussion related to this new approach:

We consider the *Configuration* and the *Orientation* of the hand as two different PPs. This approach is common to many researchers since Battison's work [2]. Initially, Stokoe [45] considered them as one, the parameter *dez*. The *Location* and *Plane* PPs are used to define the spatial position of the hands. Some authors, like Herrero [14] or Stokoe [45] merge these two PPs defining the parameter *Place* or *tab*. The Location & Plane approach has the advantage of reducing the number of units to be stored, because the system can combine the Location and Plane values automatically to obtain a spatial position. Considering the Location and the Plane as two independent PPs, we must consider the signs in which the hand contacts its Location. We introduce a specific value for the Plane PP for these situations. The Plane PP defines the horizontal distance between the hand and the body. If a contact exists, this distance is defined by the position of the Location PP value. Hence, the horizontal distance is obtained from the anatomic point defined as the Location of the sign. Muñoz includes the *Contact Point* PP to define the part of the hand that contacts its Location. Muñoz only uses this PP in the contact signs, but we have extended its usage to every sign. The Contact Point PP defines the part of the hand that must be placed at the spatial position defined by both the Location and the Plane[3]. We consider the *Movement* PP as the displacement of a hand following a defined path. We do not consider a Movement PP the transition between two consecutive Locations or Planes. Some authors define that the Movement PP includes both every change in the Configuration and Orientation, named as an *internal movement*, and every change in the position of the hands, named as an *external movement*. These differences and our reason to define the Movement

[3] For example, the hand position is different if using an all fingers extended hand shape, the palm pointing to the signer and the fingers pointing up, the Location PP is set in front of the right eye, the sign defines the Contact Point PP rather than sets the end of the little finger in front of the right eye or sets the end of the thumb in the same position.

PP as we have mentioned are exposed in the next paragraph, within the description of the Phonetic Model of this synthesizer. Finally, the Non-hand PP which groups together facial expressions and body postures. This PP consists of several independent channels corresponding to different elements of the face and the body that can move independently (e.g, the eyebrows, the eyelids, the mouth, the waist, the shoulders, etc.).

The previous paragraph described the different PPs of a sign. Next, we describe how these PPs are modified during the signing process. The phonetic model presented by Liddell and Johnson [29] describes two different blocks during the signing: the *Hold* block, when all the PPs are still, and the *Movement* block, a transition between two consecutive *Hold* blocks. The Liddell and Johnson's phonetic model description implies that the Configuration, the Orientation, the Contact Point and the Place (merge of Location and Plane) are synchronized in every Hold block. The composition of Hold and Movement blocks with their duration defines a *Segment*. Finally, a sign description requires different Segments for each hand and another segment that describes the Non-hand PP. The *Hand Tier* model [38, 39] extends the previous phonetic model, considering that the Configuration and Orientation have to be defined in their own Segment, independently of the displacement of the hands. Finally, van der Hulst and Mills [19] and Corina [4] consider the Orientation as an independent Segment. The concept of *syllable* has been also applied to SL [3, 14]. The syllable is related to the movement blocks of previous phonetic models. Brentari defined that the number of phonological movements equals the number of syllables. However, the concept of syllable itself can be described using the previous phonetic models.

After presenting the multiple phonetic models, we can describe the phonetic model that we have used in this SL synthesizer. In order to maximize the flexibility of the approach we consider that every PP is independent. Each one defines its own Hold and Movement blocks independently. These blocks can be synchronized to fulfill the Liddell and Johnson's phonetic model, or the *Hand Tier* model, but we do not implement the restrictions. The Configuration and the Orientation PPs are described using their own segments, so the previously mentioned *internal movements* correspond to the movement blocks in their own segments. The same is applied to the Location and Plane PPs, a hand displacement between two consecutive hold blocks of these PPs corresponds to their respective movement block. On the other hand, if the hand displacement describes a specific trajectory, we consider it as the Movement PP. Finally, we follow the same approach for the Non-hand PP, but considering every possible independent element of this PP (eyebrows, mouth, cheeks, shoulders, ...) within independent segments.

In order to approach SL synthesis, we must consider the different elements that can be found in a signed message. The *fingerspelling* is an alphabet representation through signs. Each letter is represented using a defined Configuration and Orientation values. Signers use fingerspelling to spell concepts or proper nouns that do not have a related sign. A *dictionary sign* (a lemma) represents a concept. It has a well known and static meaning. We can also find non *dictionary sign* used to refer to people or, some times, to neologisms. These signs are used during a conversation but they do not belong to a normative dictionary. The phonetic description of the *dictionary signs* can be modified during the signing. These flexive and inflective constructions are derived form the SL morphology, and must be handled during SL synthesis. Finally, the *classifier constructions* [8, 16, 28, 40, 41, 47] are semantically complex constructions. The approach for describing and synthesizing LSE *classifier constructions* was reported in [31]. In this work we extend the synthesis approach for the other elements.

4 Relational Database

In the previous section, we have presented the phonetic model used in this SL synthesizer. We require an approach that allows us to store the low level descriptions of every PPs, the phonetic descriptions of the signs, the relation between both of them, and every resource required for synthesizing every element present in a signed message. We also require an approach that provides enough flexibility to store the sign descriptions made using the phonetic model previously described. This model requires describing each Hold and Movement blocks for each PP of each sign. These requirements can be fulfilled using a relational database which stores the required information and establishes the required relations between different elements. The relational database also provides another advantage, the 1-to-n relations allow storing different variations of each element, this can be useful for storing in the same database dialect variations of the same concept or different mood realizations of the same PP unit. The structure of the database is presented in Fig. 1.

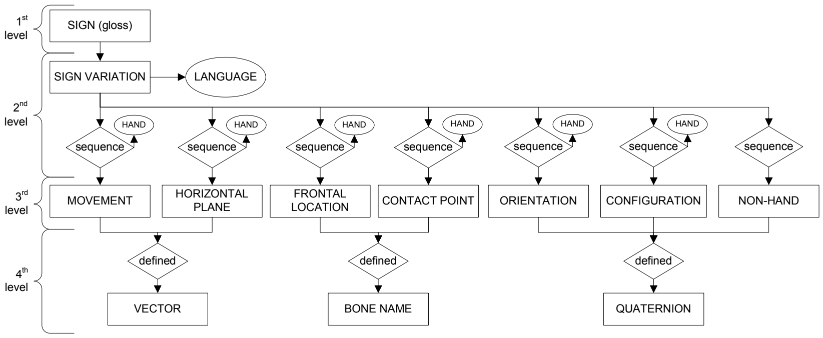


Fig. 1 Simplified entity-relationship diagram of the database

Our relational database is structured in four logical levels:

- The first level works as a dictionary entry. It consists of one table storing glosses.
- The second level corresponds to the description of a sign using sequences of PPs. Each gloss of the first level can be related to multiple entries of this level. Each entry of the 'sign variation table defines the duration of the sign in its isolated form representation. This second level contains seven tables corresponding to the seven PPs. The content of these tables is as follows:

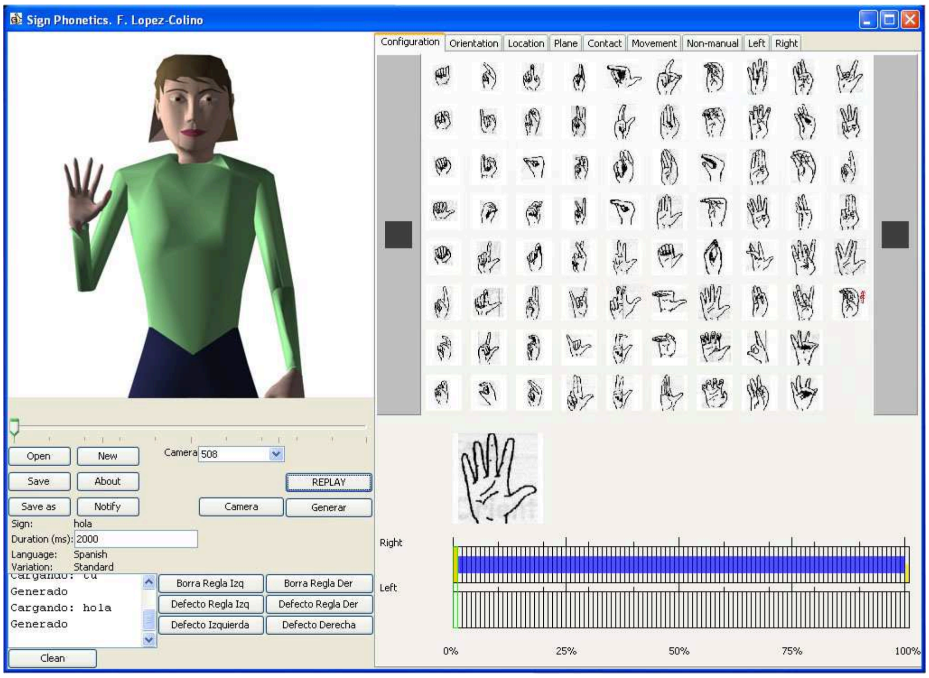
```
<PP>_sequence_table
  id_sign_variation
  id_<phoneme>
  fraction_ini
  fraction_end
  hand (only for hand related PP)
```

Each entry of these tables represents a Hold block for the corresponding PP, except for the Movement PP, which represents a Movement block. The fraction_ini and fraction_end indicate the beginning and the end of a block and they are defined using percentages of the duration of the sign. Hence, modifying the signing speed is easier. Although we consider every PP sequence independently, the synchronization proposed in different phonetic models makes possible to define the same values to the fraction_ini and fraction_end of the corresponding Hold blocks in the required PPs.

- The third level consists of a list of units for each PP. This list contains both the phonemes and the allophones for each PP [14]. Each PP also defines a "null" phoneme used for the description of partial forms.
- The fourth level is a description of each unit for every PP. The relation between the third and fourth levels of the database is 1-to-n (E.g, the same *Configuration* can be defined by different hand shapes representing different muscle tension.) We show in Fig. 1 three kinds of descriptions for the PPs; each PP only uses one of them:
  - A *Quaternion* is a common approach to define bone orientation in virtual 3D environments. The Configuration, the Non-hand and the Orientation PPs generate a relation between defined bones and a quaternion (e.g. a complete hand shape requires fifteen quaternions, one for each finger joint). The number of quaternions required to describe an expression is different for different expressions. The exact number of quaternions depends on the different bone groups, as described in Sec. 6.2, which are involved in that expression.
  - A *Bone Name* is used in two PPs, Location and Contact Point. The Location PP is described using anatomic points. Body movements preclude definitions of static coordinates for each anatomic reference. The proposed solution is to obtain required coordinates dynamically by means of skeleton bones instead of using mesh's vertexes (see Sec. 6). The Contact Point is described by means of the hand's bone used for hand positioning.
  - *Vector* is a simple 3D vector used for the Plane and the Movement PPs. When these vectors are used for the Plane PP, they define the horizontal distance to the body. In order to define a specific movement, we require movement shape and motion information, both define a sequence of position changes. These changes define hand displacement using the last position as reference, and motion information is defined as the percentages of whole movement duration for each hand displacement.

4.1 Sign descriptions in the DB

The description of a sign is stored in the second level of the database. The contents of this second level were initially based on the SignWriting descriptions and the video recordings of a LSE interpreter signing several LSE established signs. This initial set of twenty signs was used for the user recognition tests (as we will present in Sec. 9). These first signs were inserted using plain SQL commands. In order to simplify this description process, and the improvement of existing descriptions, we developed a graphic application that allows inserting new signs in the database and modifying them, using the "drag & drop" approach. The GUI is divided in two blocks (see Fig. 2), the left one contains the signing avatar and the right one allows describing the seven PPs. The right block contains one panel per PP and two additional panels to manage the descriptions of each hand (see Fig. 3). The avatar is animated using the current description made by the user, even if it is not correct. The Non-hand panel allows the definition of each independent element that composes this PP. The lower part of each panel contains two rulers, one for each hand in the hand-related panels, or several ones in the Non-hand



Camera 508

Sign: hola
Duration (ms): 2000
Language: Spanish
Variation: Standard

```
Cargando: tu
Generado
Cargando: hola
Generado
```

Fig. 2 The sign description application. This image shows the Configuration PP panel

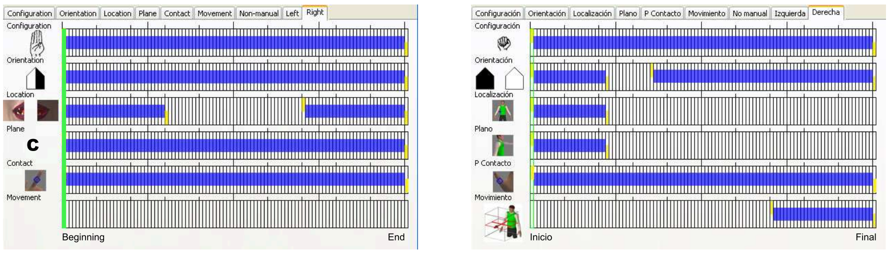
panel. These rulers represent the total duration of the sign. In order to describe a PP segment, the user drags the corresponding unit from the upper part of the panel and drops it in the corresponding ruler, defining a Hold block. The user can drag the new Hold block to its initial instant and define its duration. When defining a Movement PP unit, the process is the same but, as we mentioned before, it defines the Movement block.

This application makes the phonetic description of a sign and its modification a simple task. It allows the user to visualize the results of every modification immediately using the signing avatar. Every PP unit is described by means of images, so there is no need to learn any formal notation.

4.2 Parallel sign definitions

The database allows storing different representations of the same concept in different LSE dialects. We have stated before that a dialect variation requires a new phonetic description, so recording a different dialect, or SL definition, a new entry in the second level of the database must be done (both in the sign variation table and in the tables related to the sequences of each PP). If there is not a stored description of a dialect variation, the system tries to use the normative description of the gloss.

The mood variations mainly alter the realization of the PPs. Hence, for each unit in the third level of the database, the fourth level stores different realizations of that



(a) MOTHER

(b) SWORD

Fig. 3 Two different sign described using the application. Each bar describes a hold block for the related PP. For those PPs that are described using two phone units (like the mother's Location or the sword's Orientation), the left image corresponds to the left bar and the right image to the right one. The green bar corresponds to the precise instant of the signs' representation that the avatar shows, we have depicted both instants in Fig. 16

unit[4], related to different frames of mind. During the gesture synthesis stage (subsec. 7.1), the system automatically recovers the correct realization of the phonemes. It must be noted that current version of SiGML (HamNoSys) do not define mood variations, it only defines acceleration and tension for movements which are relevant for the sign description and its meaning. Preliminary studies have shown that the internal acceleration and duration of the different segments in a sign are mood-dependent.

4.3 Database contents

The database does not apply any restriction to its contents. Hence, we can store infinitive forms (obtained from the LSE dictionary [11]), full forms, partial definitions (lacking the definition of a PP), or templates (e.g, we can store the Configuration classifier used in a classifier constructions to refer to the object of the sentence). The database can store, for a unique gloss entry in the first level, several definitions of that gloss corresponding to an infinitive form and some other variations.

The descriptions retrieved from the database can be modified during synthesis using directives from the HLSML notation (see Sec. 5). These directives describe morphologic inflections, prosodic modifications or mood variations that determine the information that must be retrieved from the database. Hence, the database must contain infinitive forms of LSE to represent messages in LSE, as the required modifications to the signs in a sentence can be done automatically during the synthesis. However, we also stored full forms in the database for testing purposes.

The hybrid approach of the synthesizer allows both using full forms from the database (which have been previously hand tuned) and automatically synthesizing them using the directives from the HLSML notation.

[4] These realizations must not make the phoneme unrecognizable, leading to the misunderstanding of a sign.

5 HLSML: New Input Notation

The way other SL synthesizers use different XML-based notation to describe input message has been presented in Sec. 2. Both SiGML and SWML are computer friendly versions of iconographic descriptions of signs. However, using these notations requires deep knowledge of SL linguistics and some training. Our approach to sign synthesis stores phonetic definitions in a relational database (see Sec. 4), so input notation does not require to describe each sign or fingerspelling by means of their PPs. Using glosses to describe the message has been described in previous works [10, 50]. We introduce a new input notation named High Level Signing Markup Language (HLSML). This new notation uses simple tags to state a word to be spelled or the gloss of a dictionary sign. The main difference between HLSML and the gloss sequences used in other works is that HLSML also allows defining sentence's timing[5], the sign dialect[6] and the mood variation[7]; the synthesizer will automatically recover the appropriate description from the relational database and perform the required modifications. The document which describes the elements and the parameters that can be defined using the HLSML notation (HLSML's DTD) can be found at `http://www.hctlab.com/research/hci/hlsml/`.

A simple SL enunciative sentence consists of a sequence of signs. Fig. 4 presents an example of HLSML code for a simple sentence.

```
<!DOCTYPE hlsml SYSTEM "hlsml.dtd">
<hlsml>
<sentence language="lse" tag="standard">
 <sign value="one" />
 <sign value="car" />
 <sign value="red" />
 <spell value="corsa" />
</sentence>
```

Fig. 4 Example of HLSML code. This fragment defines a sentence with four units, three signs and a spelling sequence. The 'lse' value states for normalized LSE dialect. Different dialect variations for LSE are stated using 'lse-an', 'lse-ca', 'lse-va'

5.1 HLSML's phonetic level

HLSML is oriented to high level message definition. However, we consider a low level phonetic description equivalent to SiGML or SWML. The low level sign definition allows the user to describe a sign that is not stored in the relational database. Signers dynamically create signs to refer to people or concepts during a conversation. These signs are not *dictionary signs*, but they can be present during a conversation. For this reason, a low level definition has to be considered in every input notation.

[5] This includes the duration of the transition between signs (defining the duration of the Movement block between two consecutive signs) and message pauses (modifying the initial or final Hold blocks of the sign).

[6] The "language" attribute can be applied to a single sign, to a group of signs or to the sentence.

[7] The "tag" attribute defines this emotional variation. Like the "language" attribute, it can be applied to a single sign, a group of signs or a complete sentence.

In order to compare HLSML with SiGML and SWML, we present below, the required definitions to describe phonetically one sign using these three notations. The chosen sign is IR (to go), which is a two handed sign, without Non-hand PP. Figs. 5, 6(a) and 6(b) show the parametric description of this sign using HLSML, SiGML and SWML. The HLSML is quite similar to the SiGML notation but it includes the duration of the sign and it describes for each PP unit its Hold block.

```
<!DOCTYPE hlsml SYSTEM "hlsml.dtd">
<hlsml>
 <sentence language="lse">
  <signDefinition>
  <holdMoveDefinition time="1000"/>
   <configuration>
    <phoneme value="extended" side="left"
     fraction_ini="0" fraction_end="100"/>
      <phoneme value="extended" side="right"
       fraction_ini="0" fraction_end="100"/>
   </configuration>
   <orientation>
    <phoneme value="h_i_d" side="left"
     fraction_ini="0" fraction_end="100"/>
    <phoneme value="h_f_i" side="right"
     fraction_ini="0" fraction_end="100"/>
   </orientation>
   <location>
    <phoneme value="chest" side="left"
     fraction_ini="0" fraction_end="100"/>
    <phoneme value="chest" side="right"
     fraction_ini="0" fraction_end="30"/>
   </location>
   <plane>
    <phoneme value="near" side="left"
     fraction_ini="0" fraction_end="100"/>
    <phoneme value="near" side="right"
     fraction_ini="0" fraction_end="30"/>
   </plane>
   <contact>
    <phoneme value="med_end" side="left"
     fraction_ini="0" fraction_end="100"/>
    <phoneme value="med_end" side="right"
     fraction_ini="0" fraction_end="30"/>
   </contact>
   <movement>
    <phoneme value="linear_front_med" side="right"
     fraction_ini="40" fraction_end="100"/>
   </movement>
  </holdMoveDefinition>
  </signDefinition>
 </sentence>
```

Fig. 5 HLSML code for the parametric definition of the LSE sign IR (to go)

5.2 Inflective constructions

Spanish Sign Language shows, like other languages, inflective constructions which modify the phonetic description of a sign. The inflective constructions may affect different PPs of the sign: The *configuration* PP can be modified to include number information, the sign WE modifies its *configuration* PP (using the corresponding number's *configuration* PP) to represent WE-TWO, WE-THREE, etc. The same phenomenon

```
<sigml>
 <hamgestural_sign gloss="going_to_LSE">
 <sign_manual nondominant="true">
  <handconfig handshape="finger2345"
   thumbpos="halfout"
   palmor="r"
   thumbpose="in"/>
  <posture location="chest" />
 </sign_manual>
 <sign_manual >
  <handconfig handshape="finger2345"
   thumbpos="halfout"
   palmor="o"
   thumbpose="up"/>
  <posture location="stomach" />
  <motion>
   <directedmotion direction="o"/>
  <motion>
  </sign_manual>
 </hamgestural_sign>
```

(a) SiGML code

```
<swml>
 <signbox>
  <symb x="53" y="100" x-flop="0" y-flop="0">
   <category>01</category>
   <group>05</group>
   <symbnum>007</symbnum>
   <variation>01</variation>
   <fill>05</fill>
   <rotation>01</rotation>
  </symb>
  <symb x="51" y="58" x-flop="0" y-flop="0">
   <category>02</category>
   <group>05</group>
   <symbnum>001</symbnum>
   <variation>03</variation>
   <fill>01</fill>
   <rotation>01</rotation>
  </symb>
  <symb x="43" y="65" x-flop="0" y-flop="1">
   <category>01</category>
   <group>05</group>
   <symbnum>007</symbnum>
   <variation>01</variation>
   <fill>02</fill>
   <rotation>03</rotation>
  </symb>
 </signbox>
```

(b) SWGML code

Fig. 6 Parametric descriptions for the LSE sign IR (to go) in existing XML-based notations

applies to the sign HOUR: it is performed with an extended point finger aiming to the non dominant wrist (like pointing a watch) and the dominant hand performs a circular movement (like the watch's hand movement). The two hour concept is performed similar to the sign HOUR but using the hand shape of the sign TWO. The reflexive construction is performed modifying the *orientation* PP so the hands point to the signer (give vs. give me). The *location* PP can be also modified in some plural constructions; for example, to represent "three people" the signer will perform the sign PERSON but starting in three different Locations (see Fig. 7) . We have exemplified different inflective constructions used to represent number but time-, aspect- or reciprocity-related inflective constructions also modify the phonetic definition of a sign.

The inflective constructions are modifications applied to dictionary signs stored in the database. Hence, the HLSML notation describes these inflective constructions as modifiers to the <sign> using the <inflectiveModification>. These constructions require defining both the modified PP and the new value for this PP. The modified PP is defined as an attribute of <inflectiveModification>. The new phoneme for the modified PP can be defined either stating the phoneme unit that must be used (<phoneme/>) or stating a sign whose phonetic description is used for this PP (see Fig. 8).

Using phonetic-based notations like SiGML or SEA [13], it is also possible to describe a phonetic inflective construction stating the new phoneme and the resulting sign's phonetic description. The HLSML notation does not require specific phonetic knowledge to define these constructions because they may be defined stating the PP to be modified and the sign that provides the new phonetic values.

```
<!DOCTYPE hlsml SYSTEM "hlsml.dtd">
...
 <sign value="person"
  <inflectiveModification value="location">
   <phoneme value="l_shoulder" side="dominant" />
  </inflectiveModification>
 </sign>
 <sign value="person"
  <inflectiveModification value="location" side="dominant">
   <phoneme value="chest" />
  </inflectiveModification>
 </sign>
 <sign value="person"
  <inflectiveModification value="location" side="dominant">
   <phoneme value="r_shoulder" />
  </inflectiveModification>
 </sign>
```

Fig. 7 Example of an inflective construction. N-PEOPLE is signed repeating the sign PERSON n-times starting in different locations.

```
<!DOCTYPE hlsml SYSTEM "hlsml.dtd">
...
 <sign value="hour"
  <inflectiveModification value="configuration">
   <sign value="two" />
  </ inflectiveModification>
 </ sign>
```

Fig. 8 Example of an inflective construction. To represent "two hours", we declare the Configuration PP of the sign HOUR to be replaced using the configuration phonemes of the sign TWO.

5.3 Parallel behavior

Sign languages use different and independent productive organs (hands, the body and the face) defining multiple channels, two manual channels and the Non-hand channel, which is divided in several sub-channels. When the Non-hand parameter is used for prosody (not in the phonetic description of a sign), HLSML defines <nonManualSequence>. This element allows defining the duration of the Non-hand animation, if it is required. Huenerfauth's work describing Coordination and Non-Coordination [17] has been used as basis for describing the parallel actions that can occur during a signed message. These different actions can be independent, like the head shaking during a negative sentence.

The HLSML notation uses two elements for describing the sequentiality or the simultaneity of the different elements present in a signed message: The <sentence> makes that all the elements contained in this xml element are represented in sequence. On the other hand, when several elements in the signed message have to be represented at the same time in the different channels, they are contained in the same <compound> element. Both elements allow including each other so the Partition/Constitute formalism presented by Huenerfauth [17] can be represented.

6 Avatar structure

Skeleton-based animation is a common means of avatar animation. Our skeleton structure was initially defined using human anatomic model with some simplifications. At this preliminary stage of the design, the avatar skeleton was similar to the H-Anim definition and the approach of ViSiCAST.

We consider the avatar definition of these approaches can be improved in order to simplify several animation tasks: 1) Standard skeleton animation establishes that every transformation applied to a bone is automatically inherited by its descendants. This definition conflicts with the independence of the Orientation PP because a variation in the position of the hand[8] is made changing the orientation of the upper-arm and the forearm bones which will modify the orientation of the hand. 2) Face expressions are based on the mesh morphing technique, which uses an independent mesh for each expression. This approach requires to store or to transmit these mesh copies and the modifications. The creation of new expressions requires the release and distribution of the new set of expressions. 3) These approaches use the position of mesh's vertexes in order to obtain the position of an anatomic reference (e.g. the position of the chin), which implies that the mesh deformations must be calculated during the Gesture Synthesis stage consuming part of the processing resources.

We propose a new design of the signing avatar to improve these aspects: 1) the definition of the wrist has been modified inserting a new auxiliary bone. This new bone main characteristic is that it does not inherit the orientation transformations from the forearm bone. This modification ensures that the orientation of the hand will only depend on the definition of the Orientation PP. 2) Face expression use the same skeleton-based animation that is used for body animation. This approach has two advantages: it unifies the animation approach of the avatar simplifying the rendering and removes the storage and transmission requirements of the mesh morphing approach. 3) In order to obtain the position of the anatomic references defined in SL, without using the mesh, we have defined a new kind of bones, called "location bones". These new bones inherit the transformations of its parent bone the same ways as a mesh vertex. So if we require to obtain the position of an anatomic reference, we update the skeleton transformations and get the required position from the relevant "location bone" without the necessity of updating mesh deformation or even loading the mesh during the Gesture Synthesis stage. Our signing avatar, named "Yuli" is presented in Fig. 9.

The following subsections will describe the most important parts of the avatar and how the previous modifications are applied in each case.

6.1 Wrist and Hand definition

The hands are an important element in SL. Hands and wrist directly represent three of the seven PPs of sign language: Configuration, Orientation and Contact Point. They also indirectly represent the Movement, Location and Plane PPs, although these other PPs are ultimately generated by shoulder and elbow joint rotations. Thus, they have been modeled using a large percentage of total polygons. In addition, their bone structure has been especially defined to represent the independence of wrist orientation and position.

[8] The position of the hand is defined using the Location, Plane and Movement PPs.

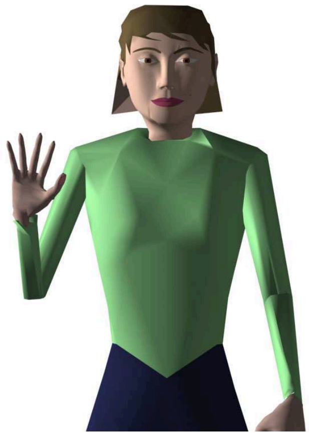


Fig. 9 Avatar mesh has been modeled focusing on face and hands. This avatar is composed by 7800 polygons

Standard skeleton structure defines a hierarchical bone structure in which transformations such as movement, rotation and scale are inherited, making it difficult to apply a specific orientation to the hand (defined by the Orientation PP). The first method is to calculate the inherited orientation from shoulder and elbow joints and compensate it; the human brain performs this process continuously. The required calculations have been eliminated simply by breaking the orientation inheritance between the first wrist bone and the forearm bone. The wrist definition establishes two bones (Fig. 10): the first has constant orientation (bone 1), whereas the second receives its orientation from the relational database (bone 2), defining the Orientation PP. This modification also simplifies the inverse kinematics process required for positioning the hand. The standard approach, which has to deal with both hand position and orientation, requires defining seven degrees of freedom (DOF). Our system manages these two properties independently so our inverse kinematics process only deals with the shoulder and elbow's four DOF.

The hand bone structure follows the design given in [32] but the Contact Point PP may require that the position of the end of the finger be calculated. This requirement is met by calculating the transformation chain inherited from the initial skeleton bone to the "location bone" at the end of the finger.

6.2 Face definition

We have defined two objectives for head and face bones. The first is to perform standard animation function. Bone transformations define mesh deformations in order to achieve the required facial expressions. The second is to act as helper elements to define the anatomic references required for the Location PP. In order to obtain those locations ,regardless of body animation, we have defined auxiliary bones. Fig. 11 shows these "location bones" such as the bones located at the ears or the forehead. These special bones do not perform mesh animation, but are used to easily obtain the position of the anatomic references without processing avatar's mesh. Therefore, during the Gesture

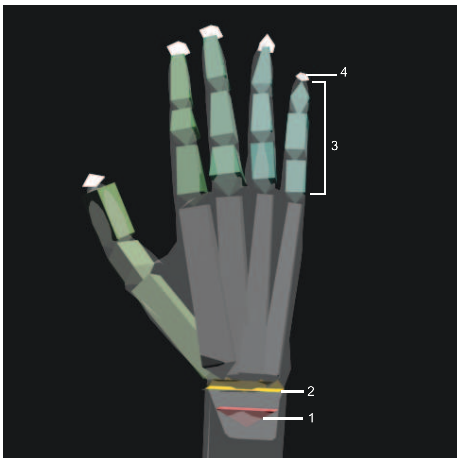
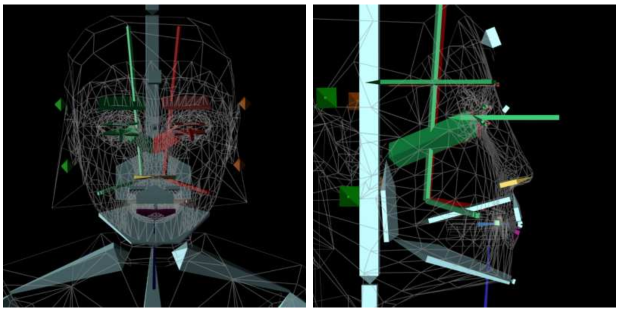


Fig. 10 Bone hierarchical structure for wrist and hand animation. The bone marked as (1) is defined as wrist-position, used for inverse kinematic process. The orientation of this bone is constant through time. The bone (2) defines the wrist-rotation, it is used for the definition of the Orientation PP. The bones identified as (3) are reserved for finger animation. At the end of the fingers we can find several "location bones" (4), these bones are used to obtain the position of the end of the fingers



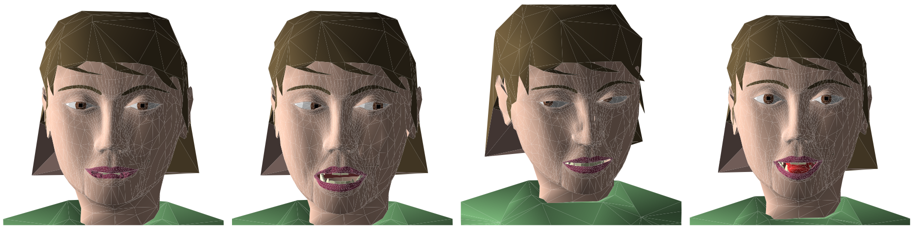
(a) Front view

(b) Side view

Fig. 11 Bone structure for face and head animation. Not every bone is used in animation process. Some of these bones, such as the one located over the head and the ones in the ears, are used as anatomic references

Synthesis process (see subsec. 7.1), the system does not require loading the mesh-related data or processing its deformation in order to obtain the position of these anatomic references.

Face expressions are composed using different bone groups. Each group represents independent parts of the face such as the eyes, the eyebrows or the mouth. Animation tracks can be assigned to each face part independently, e.g. fear and happiness expressions differ with regards to eye and eyebrow position but both have the same open mouth shape. With this strategy we can store in the database information for one mouth shape and link it with every expression that uses this mouth shape, thus reducing considerably the number of elements in the database. It also grants indepen-



(a) The mouth bites lower lip (b) The avatar looks to the left while opening the mouth (c) The avatar looks to the right while the face shows dislike (d) The avatar's tongue is out

Fig. 12 This figure shows different facial expressions obtained with the bone animation approach. The facial expression and head movements are defined in the Non-hand PP

dence to mouth shape when emulating lip movement generated by speech simulation. Fig. 12 shows some of the expressions obtained using the bone animation approach.

6.3 Body definition

Body structure was defined using two spine bones. Neck, shoulders and arm bones follow the same structure definition as does the H-Anim standard or the ViSiCAST project. Defined camera angle and position do not show legs. We only have defined thigh bones; neither mesh nor skeleton have been defined under the knees.

Sign descriptions may also require obtaining the position of different body anatomic points in order to perform signing. We have used these new special "location bones" throughout the body. These are managed the same way as head and face's "location bones", so no further explanation is required.

6.4 Avatar's mesh

Mesh modeling (Fig. 9) was done focusing on the main parts of the body used in SL, the hands and the face. It should be noted that 93.15% of mesh polygons are concentrated on the head (most of them in the face) and the hands. Equally important is the higher polygon density at frontal part of the avatar due to the camera angle. In Table 1, mesh information about polygon count is presented. The complexity of the mesh (i.e, the number of polygons) was determined based on the results of the technical validation (see Section 8).

Table 1 Polygon usage in different parts of the avatar. Main polygon concentration is located in the head and the hands because these elements need more detail than body and arms

	head	both hands	both arms	body	total
polygons	3195	4130	126	412	7863
percentage	40.63%	52.52%	1.60%	5.24%	100%

7 Distributed Architecture

The SL synthesizer has been designed to accommodate a great diversity of final user devices. In order to cover most hardware and software platforms, a distributed architecture has been established, separating the whole process into three steps: Gesture Synthesis, Rendering and Visualization (see Fig. 13).

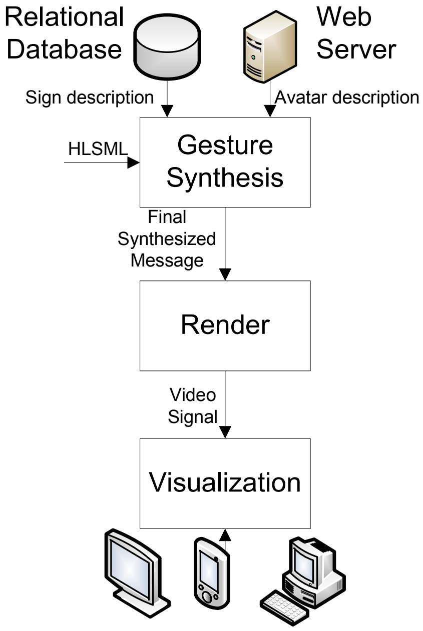


Fig. 13 These are the main modules of the SL synthesizer. The defined communication protocol allows each element to be executed independently. Hence, the synthesizer can be adapted to many devices with different resources

7.1 Gesture Synthesis

The first stage of the SL synthesis process is to generate the animation tracks corresponding to the received sign message. The gesture synthesis module (see Fig. 14) receives an HLSML message. The HLSML describes a SL message. Also, it may contain modifiers to the message, parameters such as whole phrase speed, single sign speed or movement stress, mood variation and dialect variation (see subsec. 5).

The next step during the gesture synthesis process is to obtain the avatar's description. This module downloads a 'm3g' file[9] containing all required elements, such as cameras, lights, materials, avatar's mesh and skeleton structure. This file does not include textures, but they can be downloaded, if needed.

In order to create the animation tracks for each bone, the synthesizer must obtain each sign description from the relational database. The database contains normalized

[9] This file format has been established by the JSR-184 standard.

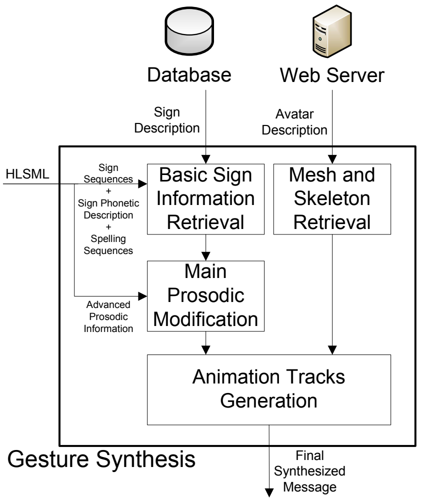

Database
Web Server
Sign Description
Avatar Description
HLSML
Sign Sequences + Sign Phonetic Description + Spelling Sequences
Basic Sign Information Retrieval
Mesh and Skeleton Retrieval
Main Prosodic Modification
Advanced Prosodic Information
Animation Tracks Generation
Gesture Synthesis
Final Synthesized Message

Fig. 14 Detailed diagram of the gesture synthesis module.

sign descriptions (infinitive forms), full forms, partial forms and templates. The HLSML message can include information to modify these descriptions (inflections), the missing information required to complete the templates or the whole phonetic description of a sign (subsec. 5.1). Depending on the attributes included in the HLSML message, this module modifies the queries to the database in order to retrieve the correct information from it (see Sec. 4). The final animation tracks are created based on the information stored in the database and input message descriptions and modifiers.

The hand shape and the hand orientation influence wrist final position, inverse kinematic calculations to define shoulder and elbow rotations are delayed to the end of the synthesis. An advantage of using the new bone in the wrist (see subsec. 6.1) is that the hand orientation is independent of its position. Hence, the inverse kinematics algorithm can be simplified omitting the wrist's degrees of freedom and just focus on the shoulder's and elbow's degrees of freedom, which are just four. In this first version of the sign language synthesizer, a simple iterative inverse kinematics algorithm has been used. Once an animation track is defined, it is assigned to the corresponding bone and related to a global timing element that provides simultaneity between different animation tracks.

7.2 Rendering

Two rendering strategies must be defined based on the visualization process. The first one is focused on real time visualization. The main objective is to achieve an optimal frame rate in order to provide a fluid animation. The program calculates the corresponding rendering instants based on the last frame processing time. This time depends on many factors such as scene complexity, the number of mesh polygons, active animation tracks and hardware and software resources.

Frame rendering duration has influence on real time visualization but, if the presentation process can be delayed, frame rendering duration has no effect. The main task of this option is to create animated sequences of the generated sign message. These sequences can be stored for further use in a video file. However, video settings should be defined using different settings for size, frame rate, color depth and video compression, according to demand.

7.3 Visualization

The visualization of the resulting sequence is the last stage of the process, which has two different possibilities. The first one involves performing the visualization directly from rendering output. This option can only be used if the client device has sufficient graphic resources for the rendering process. The second option is reserved for devices with low 3D capabilities or without the required rendering API. In this case, the visualization process will consist of playing a video that will be downloaded after the server has finished rendering the whole message or while the message is being rendered using streaming technology.

7.4 Adaptation to user device

Three different and independent stages have been defined above for the whole sign synthesis process. Each stage can be assigned either to the final client device or to a synthesizer server, except for visualization that must be run on the client side. The distribution of these three stages between server and client side defines several different scenarios. Each scenario suits different client, server and network resources[10], so multiple user device adaptation is possible. More than one scenario may sometimes be available. A specific module must optimize server and network load. This element will choose the optimal solution for each session depending on network and server load and the resources of the client's device. A deeper discussion of this section can be found in [30].

8 Validation process

The whole synthesis system was tested on a Pentium IV, 2GHz with 512 MB of RAM memory and 8 MB of video memory. The operating system was Windows XP SP2 and we used Hybrid Rasteroid 3 [20], a Windows implementation of the JSR-184 API for J2SE.

The validation tests were performed using the desktop implementation of the synthesizer. The aim of the test was to ensure a high enough frame rate to obtain fluid animation on a computer. Fig. 15 shows a rate of twenty images per second, a fps rate within the limits of a fluid animation. In order to obtain this rate, several mesh optimizations were performed, such as the reduction of the number of polygons and the creation of smoothing groups.

[10] We have only considered client-server communication.

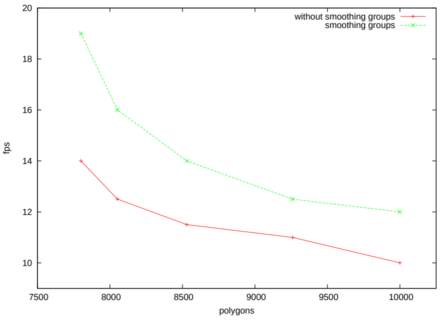


Fig. 15 Frame rate related to the number of mesh's polygons. The importance of using smoothing groups must be emphasized, as doing so raises the fps rate significantly

9 Results and Evaluation

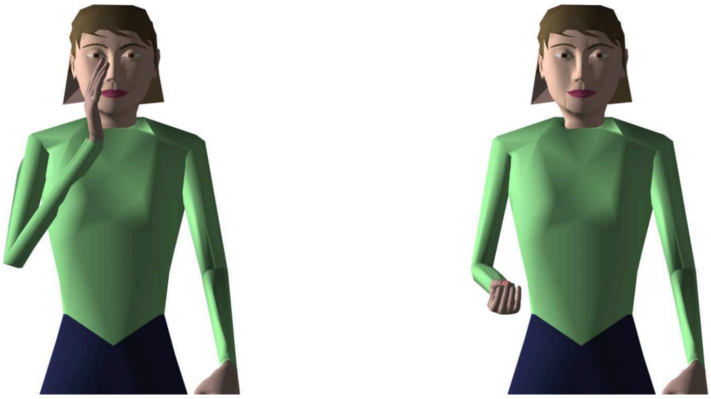
The previous test only provided information about animation fluidness and verified the correctness of the implementation. Obviously, every synthesizer should be evaluated using native evaluators. Three different sets of experiments were performed with LSE signers to check message understanding.

9.1 Experimental setup

In the first set of experiments, a group of twenty signs were introduced in the database. These signs were: HELLO, I, DEAF, GROW_UP, HERE, Madrid, AGE, 25, HAPPY, PARTNER, SCHOOL, Toledo, HOUR, NEAR, MINE, GREEN (color), RED, TOMORROW, ALL_DAY, TODAY. The signs were chosen to be representative of all kinds of signs (single- and double-handed, with and without non-hand PP). The descriptions of these signs were obtained from a paper dictionary which related the gloss to a SignWriting description; these descriptions were complemented with video recordings of a deaf person signing each of them, which provided timing information. This task was performed by a computer expert with only theoretical knowledge of LSE. To facilitate the final testing, the set of signs were rendered into a 20 fps video (Fig. 16). This video was presented to a group of six LSE experts, who were three hearing interpreters of LSE and three deaf LSE natives. These six experts work as teachers at the same LSE academy. Each sign was presented once and the users had to identify each one before viewing the next one. The deaf users communicated to an interpreter if they had recognized the sign and, if it was so, which sign did they recognized. The same procedure was repeated in the three evaluations.

9.2 First evaluation

The results of this first test are presented in Table 2. The obtained recognition rates were as follows: 77% of recognition rate among hearing teachers and 58% among deaf teachers. The results were promising if the way the signs were defined is considered, but these results are significantly lower than the results obtained using other synthesizers.



(a) MOTHER (b) SWORD

Fig. 16 Two different frames of the signing avatar. These frames correspond to the descriptions depicted in Fig. 3

However, they provided very important and useful suggestions about avatar's appearance and performance. They also showed that deaf subjects are more demanding when evaluating sign synthesis, so the evaluation must succeed on deaf subjects.

Table 2 Results of the first recognition test. Users 1-3 were deaf teachers and 4-6 were hearing teachers

	User group	
	1-3	4-6
Recognition rate	58%	77%
Average recognition rate	67.5%	

9.3 Improving the sign descriptions

The previous conclusion supports the main motivation for developing our approach: deaf people must have an active role in the definition and tuning of the signs. The synthesizer must be enough flexible and precise to incorporate all the suggestions and the modifications proposed by the LSE signers, especially those suggested by the LSE natives.

The definitions of the signs were improved using all the suggestions provided by the experts. The collaborative work consisted in a session where an expert visualized a sign, proposed several modifications to the definition of several PPs units, most of them hand-shapes. These modifications also included altering the temporal aspects of the sign, such as "this hand-shape should remain still a bit more" or "the transition between these two orientations should be faster than the transition between these two locations". These adjustments are related to the Hold and Movement blocks of the phonetic model. It must be noted that these modifications cannot be done using SiGML or SWML notations. The checking was applied to every sign of the first test even if it was correctly recognized. Obviously, the signs that were identified in the first test required fewer modifications.

9.4 Second evaluation and recognition results

For the second test, we included two new users in the testing group. These new users were LSE natives, who were used as reference, as they had no previous experience with the avatar. The same signs were presented to the eight people, in different order from the first time, and they were asked to identify each sign. The users of the first test were not informed about the correctness of their previous answers. Therefore, if a sign was not recognized in the first test and was not improved enough it would not be correctly identified in this second test. The results of this second test are presented in Table 3. We have obtained an average recognition rate of 82.3%, compared to the 81% of recognition rate obtained by the ViSiCAST (see subsec. 2.1 for more information about their experimental setup); it shows that allowing deaf people to introduce modifications to the definition and temporal evolution of a sign increases the quality of the synthesized signed messages.

Table 3 Results of the second recognition test. Users one to six are the same ones from the first recognition test and users seven and eight are the new users introduced for this experiment. The final average recognition rate is the recognition average of each group, instead of user's average

	User group		
	1-3	4-6	7-8
Recognition rate	82%	90%	75%
Average recognition rate	82.3%		

The results of this second test show that the recognition rate reported by the new users (7–8) was lower than the recognition rate presented by the other deaf users (1–3) in the same experiment. This difference was expected as this second experiment was the first contact of these users with the signing avatar. However, when comparing the first-time results of both groups (users 1–3 in the first test and users 7–8 in the second test) an increase of the recognition rate can be observed, from 58% to 75% of correct answers. This increase in the recognition rate during the first-time evaluation is a consequence of the sign description tuning performed by the LSE expert.

9.5 Third evaluation, using the PP application

During the last evaluation, we proposed another set of twenty different signs to the same group of eight signers. This time, the signs were inserted in the database using the application we have presented in subsec. 4.1. In order to test the usability of the application, we asked two students of Computer Science, with no previous knowledge of LSE to perform the signs' descriptions. We showed them how to use the application and some basic notions of LSE phonology. Each of them inserted ten signs in the database using as reference a LSE video dictionary [11]. The selected signs for this third test were: MOTHER, SWORD, WATER, COAT, FINISH, TELEVISION, ROAD, DUSK, HOUSE, BUILDING, CAR, CHURCH, BROTHER, BLUE, SUBWAY, ORANGE (fruit), DOOR, TO_MEET, TELESCOPE, TO_WORK.

The obtained results of this third evaluation are shown in Table 4. We can observe the increase of the recognition rate of the third group of users, related to their adaptation to the avatar's signing style. The average recognition rates of the second and third evaluations are quite similar, but the description of the signs used for the third evaluation was not verified by SL experts. Although the recognition rates could be improved if this verification was to be performed. It can be observed that the recognition rates reach an upper limit, which depends on the avatar's quality, the animation performance, etc.

Table 4 Results of the third recognition test. The groups and the average calculation approach is the same one as in the previous test

	User group		
	1-3	4-6	7-8
Recognition rate	82%	88%	80%
Average recognition rate	83.3%		

10 Conclusion

This new approach to sign language synthesis presents a hybrid paradigm; it improves the phonologic descriptions of signs allowing manual modifications. These modifications include the timing of the sign description, allowing the modification of the Hold and Movement blocks of each PP independently. This is possible due to the flexibility of the proposed phonologic model.

The database also allows storing different realizations of a sign, all of them related to each other by means of a common entry in the database. The database can store infinitive forms of a sign, prosodic modified forms, partial forms or templates used for other elements in a message. The database can also store different dialect realizations of the same gloss. In this approach the PPs descriptions are stored in a database, which releases the input notation of this description task, allowing to simplify the input notation and to focus on message description.

HLSML is a new xml-based notation which extends the possibilities of existing gloss based approaches, allowing the description SL sentences composed by fingerspelling sequences, dictionary signs and non dictionary signs. These elements can be altered by prosody or inflective modifiers. The inflective modifications can be defined using the glosses of the modified and modifier signs, so it is not necessary to know the name of the PP units. This notation also allows defining the multiple and concurrent channels that describe a signed message.

The avatar's design improves previous definitions to simplify the gesture synthesis process. These new characteristics mean: 1) the unification of body and face animation approach. 2) The independent management of all PPs (specially the Hand Orientation PP). 3) The simplification of the inverse kinematics algorithm as it only requires handling four degrees of freedom instead of seven. 4) The avoidance of mesh deformations management during the Gesture Synthesis process.

The obtained results during the user evaluations show an increase of the sign recognition rate in the first encounter with the avatar after hand-tuning the temporal aspects of the PPs sequences. This initial rate should be considered as a base line as, like other works have shown, the recognition rate raises when the users get used to avatar's signing style. We have also presented that using the sign description application we have developed, people with no knowledge of SL's phonology can easily describe the signs by means of their PPs. These descriptions can be used for SL synthesis obtaining acceptable sign recognition rates, similar to other synthesis works.

11 Future work

Future work will deal with the integration of this LSE synthesis module and the LSE machine translation synthesis module we are developing and a speech recognition system, to obtain a full Spanish-to-LSE machine translation system. Although the system is capable of synthesizing full sentences, there is still much work to do. Improving the synthetic messaes will require the automatic adaptation of the stored phonetic descriptions to continuous synthetic signing: the transition between signs and signing pauses modify the initial and final Hold blocks of the definition. Including sentence prosody will modify the speed of the different Hold and Movement blocks of the signs, and include inflective constructions related to the Non-hand PP. It has been also observed the different realizations of the phonemes depending on the previous signs (allophones).

The SEA notation [13] has been used as the phonetic notation in the LSE normative dictionary. This notation is based on a syllabic phonologic model, we are considering the development of an application that inserts automatically a first phonetic description in the database using these SEA strings. This first version can be later modified and enhanced using the developed application.

During the evaluations we compile the opinions of all users regarding avatar's look. Although the avatar's appearance and details were enough to correctly distinguish the hand shapes and face expressions, users expressed that avatar's appearance should be more realistic. We will develop a new human-like avatar. This new avatar will incorporate the required resources to avoid the collision of the hands and the body that we have observed in few signs.

Acknowledgements Authors would like to acknowledge the teachers of the SIGNAR academy for their participation on the evaluations and to the FPU-UAM program for its financial support. Authors are grateful to journal's editor and reviewers for their comments and suggestions which have improved this paper and proposed interesting future lines of research.

References

1. Bangham A, Cox S, Elliot R, Glauert J, Marshall I (2000) Virtual signing: Capture, animation, storage and transmission - an overview of the visicast project. In: IEE Seminary on Speech and Language Processing for Disabled and Elderly People
2. Battison R (1973) Phonology in american sign language: 3-d and digit-vision. In: Proceedings of the California Linguistic Association Conference
3. Brentari D (1996) Trilled movement: Phonetic realization and formal representation. Lingua 98(1–3):43–71

4. Corina DP (1996) Sign linguistics phonetics, phonology and morpho-syntax. Lingua 98(1–3):73 –102
5. Cox S, Lincoln M, Nakisa M, Wells M, Tutt M, Abbott S (2003) The development and evaluation of a speech-to-sign translation system to assist transactions. International Journal of Human-Computer Interaction 16(2):141–161, DOI 10.1207/S15327590IJHC1602_02
6. Elliot R, Glauert J, Jennings V, Kennaway R (2004) An overview of the sigml notation and sigml signing software system. In: Proceedings of Language Resources and Evaluation Conference, Lisbon, pp 98–104
7. Elliott R, Glauert J, Kennaway R, Marshal I, Sáfár É (2008) Linguistic modelling and language-processing technologies for avatar-based sign language presentation. Universal Access in the Information Society 6(4):375–391, DOI http://dx.doi.org/10.1007/s10209-007-0102-z
8. Emmorey K, Herzig M (2003) Perspectives on Classifier Constructions in Sign Languages, Psichology Press, chap 10: Categorical Versus Gradient Properties of Classifier Constructions in ASL, pp 221–246
9. European Media Masters of Art (2002) Vsign. http://www.vsign.nl/
10. Fotinea SE, Efthimiou E, Caridakis G, Karpouzis K (2008) A knowledge-based sign synthesis architecture. Universal Access in the Information Society 6(4):405–418
11. Fundación CNSE (2008) Diccionario Normativo de la Lengua de Signos Española [Spanish Sign Language Normative Dictionary]. Fundación CNSE
12. Hanke T (2004) Hamnosys - representing sign language data in language resources and language processing contexts. In: Heidelberg SB (ed) Proceedings of LREC, Lisbon
13. Herrero Blanco Á (2004) A practical writing of sign languages. In: Proceedings of LREC, Lisbon, pp 37–42
14. Herrero Blanco A (2009) Gramática Didáctica de la Lengua de Signos Española (LSE), 1st edn. Ediciones SM
15. Huang Z, Eliëns A, Visser C (2002) Step: A scripting language for embodied agents. In: Proceedings of the Workshop on Lifelike Animated Agents, Tokyo, pp 46–51
16. Huenerfauth M (2004) Spatial representation of classifier predicates for machine translation into american sign language. In: Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation
17. Huenerfauth M (2005) Representing coordination and non-coordination in an american sign language animation. In: Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility, ACM, New York, NY, USA, pp 44–51, DOI http://doi.acm.org/10.1145/1090785.1090796
18. Huenerfauth M (2009) A linguistically motivated model for speed and pausing in animations of american sign language. ACM Transactions on Accessible Computing 2(2):1–31, DOI http://doi.acm.org/10.1145/1530064.1530067
19. van der Hulst H, Mills A (1996) Issues in sign linguistic: Phonetics, phonology and morpho-syntax. Lingua 98(1–3):3 – 17
20. Hybrid (2006) Hybrid rasteroid. http://www.hybrid.fi/
21. Igi S, Ujitani M, Tamaru M, Yamamoto Y, Sugita S (2003) Sign-language synthesis for mobile environments. In: Proceedings of The 11 International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, URL `http://wscg.zcu.cz/wscg2003/Papers_2003/C19.pdf`

22. Informatics and Telematics Institute (2004) Vsigns. http://vsigns.iti.gr:8080/VSigns
23. ISO/IEC 14772-1:1997 (1997) Information technology – Computer graphics and image processing – The Virtual Reality Modeling Language (VRML) – Part 1: Functional specification and UTF-8 encoding. International Organization for Standardization, Geneva, Switzerland
24. ISO/IEC 14772-2:2004 (2004) Information technology – Computer graphics and image processing – The Virtual Reality Modeling Language (VRML) – Part 2: External authoring interface (EAI). International Organization for Standardization, Geneva, Switzerland
25. ISO/IEC 19774:2005 (2005) Information technology – Computer graphics and image processing – Humanoid animation (H-Anim). International Organization for Standardization, Geneva, Switzerland
26. Karpouzis K, Caridakis G, Fotinea SE, Efthimiou E (2007) Educational resources and implementation of a greek sign language synthesis architecture. Computers & Education 49(1):54 – 74, DOI DOI:10.1016/j.compedu.2005.06.004, URL http://www.sciencedirect.com/science/article/B6VCJ-4H4V6WF-1/2/7485593c58f51483c6da1d049922c501
27. Kennaway R, Glauert J, Zwitserlood I (2007) Providing signed content on the internet by synthesized animation. ACM Transactions on Computer-Human Interaction 14(15):1–29, DOI http://doi.acm.org/10.1145/1279700.1279705
28. Liddell SK (2003) Perspectives on Classifier Constructions in Sign Languages, Psichology Press, chap 9: Sources of Meaning in ASL Classifier Predicates, pp 199–220
29. Liddell SK, Johnson RE (1989) American sign language: The phonological base. Sign Language Studies 64:195–278
30. López-Colino F, Colás J (In Press) Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts, IGI International, chap Providing ubiquitous access to synthetic sign language contents over multiple platforms
31. López-Colino F, Garrido J, Colás J (2009) Description and synthesis of spanish sign language classifiers. In: Latorre P (ed) Proceedings of INTERACCION, X International Conference of Human-Computer Interaction, AIPO, Barcelona, Spain
32. Moccozet L, Magnenat-Thalmann N (1997) Dirichelet free_form deformations and their application to hand simulation. In: Proceedings of Computer Animation'97, IEEE, pp 93–102
33. Muñoz Baell I (1999) Cómo se articula la lengua de signos española? Confederación Nacional de Sordos de España
34. Prillwitz S, Leven R, Zienert H, Hanke T, Herming J (1989) HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide. Signum-Verlag
35. Rocha A, Pereira G (2004) Supporting deaf sign languages in written form on the web. In: 4th International Conference on Language Resources and Evaluation, Association pour le Traitement Automatique des Langues, TALN, pp 26–28
36. San Segundo R, Barra R, D'Haro LF, Montero JM, Córdoba R, Ferreiros J (2006) A spanish speech to sign language translation system for assisting deaf-mute people. In: Proceedings INTERSPEECH-2006, Interspeech, Pittsburgh. USA, pp 1399–1402
37. San Segundo R, Barra R, Córdoba R, D'Haro LF, Fernández F, Ferreiros J, Lucas JM, Macías-Guarasa J, Montero JM, Pardo JM (2008) Speech to sign language

translation system for spanish. Speech Communication 50(11-12):1009–1020, DOI http://dx.doi.org/10.1016/j.specom.2008.02.001

38. Sandler W (1989) Phonological Representation of the Sign: Linearity and Nonlinearity in American sign Language. Foris Publications Holland
39. Sandler W, Lillo-Martin D (2006) Sign Language and Linguistic Universals. Cambridge University Press
40. Schembri A (2003) Perspectives on Classifier Constructions in Sign Languages, Psichology Press, chap 1: Rethinking 'classifiers' in Signed Languages, pp 3–34
41. Schembri A, Jones C, Burnham D (2005) Comparing action gestures and classifier verbs of motion: Evidence from australian sign language, taiwan sign language, and nonsigners' gestures without speech. Journal of Deaf Studies and Deaf Education 10(3):272–290
42. Segouat J (2009) A study of sign language coarticulation. In: ACM SIGACCESS conference on Computers and Accessibility, ACM, New York, NY, USA, 93, pp 31–38, DOI http://doi.acm.org/10.1145/1531930.1531935
43. Segouat J, Braffort A (2009) Toward the study of sign language coarticulation: Methodology proposal. In: Proceedings of the International Conference on Advances in Computer-Human Interaction, IEEE Computer Society, Los Alamitos, CA, USA, pp 369–374, DOI http://doi.ieeecomputersociety.org/10.1109/ACHI.2009.25
44. Solina F, Krapez S, Jaklic A, Komac V (2001) Design and Management of Multimedia Information Systems, Idea Group Publishing, chap 13: Multimedia Dictionary and Synthesis of Sign Language, pp 268–281
45. Stokoe WC (1960) Sign language structure: An outline of the visual communication systems of the american deaf. Studies in linguistics: Occasional papers 8:1–78
46. Stokoe WC (1978) Sign Language Structure: the first linguistic analysis of American sign language. Linstok Press Incorporated
47. Supalla T (2003) Perspectives on Classifier Constructions in Sign Languages, Psichology Press, chap 11: Revisiting Visual Analogy in ASL Classifier Predicates, pp 249–257
48. Sutton V (1974) Signwriting. http://www.signwriting.org/
49. VCom3D (2009) Signsmith studio. Online (www.vcom3d.com), URL `http://www.vcom3d.com/signsmith.php`
50. Zhao L, Kipper K, Schuler W, Vogler C, Badler NI, Palmer M (2000) A machine translation system from english to american sign language. In: AMTA '00: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future, Springer-Verlag, London, UK, pp 54–67
51. van Zijl L (2006) South african sign language machine translation project. In: ACM SIGACCESS conference on Computers and Accessibility, ACM, New York, NY, USA, pp 233–234, DOI http://doi.acm.org/10.1145/1168987.1169031
52. Zwiterslood I, Verlinden M, Ros J, Schoot S (2004) Synthetic signing for the deaf: esign. In: Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment, CVHI, Granada, Spain