Motion synthesis and editing for the generation of new sign language content

Lucie Naert, Caroline Larboulette, Sylvie Gibet

Machine Translation manuscript No.
(will be inserted by the editor)

Motion Synthesis and Editing for the Generation of New Sign Language Content

Building New Signs With Phonological Recombination

Lucie Naert · Caroline Larboulette · Sylvie Gibet

Received: date / Accepted: date

Abstract Existing work on the animation of signing avatars often relies on pure procedural techniques or on the playback of Motion Capture (*MoCap*) data. While the first solution results in robotic and unnatural motions, the second one is very limited in the number of signs that it can produce. In this paper, we propose to implement data-driven motion synthesis techniques to increase the variety of Sign Language (SL) motions that can be made from a limited database. In order to generate new signs and inflection mechanisms based on an annotated French Sign Language *MoCap* corpus, we rely on phonological recombination, i.e. on the motion retrieval and modular reconstruction of SL content at a phonological level with a particular focus on three phonological components of SL: hand placement, hand configuration and hand movement. We propose to modify the values taken by those components in different signs to create their inflected version or completely new signs by (i) applying motion retrieval at a phonological level to exchange the value of one component without any modification, (ii) editing the retrieved data with different operators, or, (iii) using conventional motion generation techniques such as interpolation or inverse kinematics, which are parameterized to comply to the kinematic properties of real motion observed in the data set. The quality of the synthesized motions is perceptually assessed through two distinct evaluations that involved 75 and 53 participants respectively.

Keywords Sign Language · Motion Synthesis · Motion Capture · Avatar · Phonological Recombination

L. Naert
IRISA, Université Bretagne Sud
E-mail: lucie.naert@univ-ubs.fr

C. Larboulette
IRISA, Université Bretagne Sud
E-mail: caroline.larboulette@univ-ubs.fr

S. Gibet
IRISA, Université Bretagne Sud
E-mail: sylvie.gibet@univ-ubs.fr

1 Introduction

Sign languages (SL) are the natural languages of deaf people. In SL, the movements of the whole body are used to convey a message that is interpreted by the interlocutor through his or her visual channel. While video recordings lack the depth information of human motion, and their editing, storage and analysis are complex issues, signing avatars, which are virtual characters communicating in SL, are a promising alternative technology. Indeed, they make it possible to preserve the anonymity of the signer and to gain interactivity: the appearance, signing speed and point of view can be altered to meet the needs of the user. However, for this technology to be competitive, the avatar must be fully animated with precise, realistic and meaningful motions because it must be understood and accepted by the deaf community.

Existing work on signing avatar animation either relies on pure procedural techniques or on data-driven techniques. While the first solution results in precise, unlimited but robotic motions, the second one produces realistic animations but the number and variety of the signs that can be created in this way are limited by the original database.

In this paper, we describe and implement data-driven motion synthesis techniques to increase the variety of signs that can be built from a limited database. To generate new signs based on an annotated *MoCap* corpus, we rely on **phonological recombination**, i.e. on motion retrieval and modular reconstruction of SL content at a phonological level. We use the *MoCap* data both as synthesis material that can be played back and edited, and as analysis material to parameterize existing procedural synthesis techniques so that the resulting animations are as close to reality as possible. The results of the procedural techniques complement the *MoCap* movements when motion playback and editing are not available, or when some processing is necessary due to a change in the context of the signed utterance.

In this work, we rely on a skeletal animation system developed to control a skeleton from captured data (**?**). Our initial *MoCap* database, called *LSF-ANIMAL* (**?**), contains one hour of raw French Sign Language (LSF) data annotated on different parallel tracks at a gloss and at a phonological level.

The remainder of this paper is organized as follows: Section 2 gives some background on the linguistic context of SL and reviews existing work on signing avatars. Section 3 details our approach and defines our phonological recombination technique. The three following sections present the synthesis methods developed to create new linguistic content organized according to the phonological component involved, respectively the hand placement component (Section 4), the hand configuration (Section 5) and the hand movement (Section 6). Section 7 presents the results of the qualitative evaluations of the proposed techniques. Finally, Section 8 concludes and presents perspectives for future work.

2 Related Work

The work presented in this paper is at the intersection of two domains: linguistics of sign languages and motion synthesis.

2.1 Linguistic Background of Sign Languages

The **phonological** (or **parametric**) approach states that a sign is a sequence of discrete values taken, in parallel, by five phonological components of SL (**???**):

1. The **hand configuration** corresponds to the overall shape of the hand characterized by the posture of the fingers. Each configuration corresponds to a discriminating and meaningful hand shape.
2. The **hand placement** is the location of the hand in the signing space or on the body of the signer. For the computer animation community, it designates the discrete area or the specific coordinates where the hand is positioned at a precise time.
3. The **hand movement** represents the trajectory of the wrist over time. Unlike hand configuration which takes a value in finite sets, hand movement is continuous and can represent any trajectory.
4. The **hand orientation** is defined by the direction of the hand palm and of the palm normal.
5. The **non-manual features** (NMFs) include the facial expressions, the mouthing, the gaze, and torso direction.

The phonological approach thus states that a sign is a sequence of values taken in parallel by each of these components in finite sets (hand configuration) or infinite sets (hand movement, hand orientation). However, while in oral languages, words are formed with a simple sequence of phonemes, in sign languages, the components take on values simultaneously in addition to having a sequential aspect. To designate this particularity, we refer to sign languages as multilinear languages (**?**).

Furthermore, a sign can be either in its **citation form** (**???**) or in an **inflected form**. The citation form is the default form of the sign deprived of any syntactic context. It is present in dictionaries or educational applications teaching isolated signs. In order to take into account the contextual information of an utterance, the transformation of the citation form of a given sign results in an inflected form of that sign. For example, the amplitude of the motion of a sign designating an entity can be modified to describe the size of the entity. Translation or storytelling applications require the generation of full utterances with inflection mechanisms. In this paper we focus on the synthesis of signs, whether in their citation or inflected form, through phonological recombination.

2.2 Sign Synthesis

Isolated signs can be created using hand-crafted or automatically computed keyframes, procedural animation, data-driven techniques, or hybrid techniques.

Keyframe-based techniques, either hand-crafted (**??**) or automatically generated (**??**), give a non-continuous definition of motion where each keyframe is a given posture of an avatar at a given time. As the number of keyframes is too small to define a smooth motion, interpolation between two consecutive keyframes must be calculated. The resulting motion greatly depends on the definition of the keyframes and on the complexity of the interpolation. Hand-crafted animation is a tedious process for which the realism of the results depends on the skills and choices of the graphics designer. Automated keyframing animation is based on

keyframes generated using isolated targets and forward and inverse kinematics algorithms.

Procedural techniques take advantage of the temporal control of systems (whether kinematic or dynamic), using cost functions to be minimized to achieve objectives (e.g., reaching targets), in order to create continuous motion (**??**). Both automatic keyframing and procedural techniques can be coupled with a high-level representation of the sign that allows a phonological construction of the sign. For these methods, the granularity of the specification level can facilitate precise, configurable and flexible animation, but may produce unnatural and jerky movements due to the chosen optimization process, and moreover does not ensure synchronization between phonological components.

Those synthetic synthesis methods lack the naturalness and expressive qualities of human motion whereas, in **data-driven techniques** (**???**), the resulting animation has the authenticity of natural human motion without the need for additional treatments. However, it is difficult to generate new sign language content from a limited set of movements in the database, and contextual variations in the captured motions are not easily synthesized, which is a problem given the high iconicity and variability of sign languages. Machine learning methods (**?**) are a promising way to create new content from a limited *MoCap* database, although, as data-driven techniques, they will still produce utterances influenced by the input data.

Hybrid methods take advantage of the temporal accuracy of keyframing techniques, the automation of procedural techniques, and the naturalness of data-driven methods. Some of these approaches have emerged in recent years. For example, the work of Lombardo et al. combines keyframing with *MoCap* data (**??**) while the *Paula* animation system mainly relies on hand-crafted keyframes processed by the Azee specification language to build complex SL mechanisms, for example to synthesize proforms (**?**), combined with *MoCap* data (**???**).

We propose to take advantage of the realism of data-driven approaches and the flexibility of procedural methods with **phonological recombination** in order to create new SL content. More precisely, following SL linguistic works, we propose to enrich our initial database by taking advantage of a phonological definition of LSF. We consider three components: hand placement, hand configuration and hand movement. We only focus on the manual characteristics of the signs, as gaze direction, facial expression and body movements require other animation methods.

3 Phonological Recombination

We call **phonological recombination** the act of modifying each phonological component independently in order to create new content and, in this way, to enrich the initial database. We noticed indeed that new signs and inflections could be created by modifying the value taken by only one of the phonological components independently of the others (e.g., changing the hand placement of a sign enables to create its spatialized form). Table 1 lists the SL mechanisms we are working on in this paper with respect to the selected phonological components.

Considering a database annotated at a phonological level, it is possible to retrieve and modify the value taken by one phonological component during the realization of a sign and to overwrite the existing value for the component with this

Table 1 List of different SL mechanisms in relation to the selected phonological components (inspired from (**?**)). The mechanisms in parenthesis are perspectives for future work.

Level	Hand Placement	Hand Configuration	Hand Movement
Phonological		Dactylology	Articulatory motion
Sub-lexical		Derivative base	
Lexical	Spatialization	Shape Specifiers	(Size and shape specifiers)
Syntactic	Pointing	(Proform)	(Trajectories)

new value to create a new sign. This technique has the advantage of enabling the creation of new realistic content as it is based on human data, however, in practice, a motion retrieval/overwriting operation alone is not enough to obtain correct content. Several challenges must be addressed in order to achieve an acceptable result:

1. If the value of the component with which we want to overwrite the current value does not exist in the database, it is necessary to create it. To do this, it is possible to modify existing values or to synthesize them from scratch using existing procedural methods.
2. The modification of a channel[1] can have an impact on joints that are not part of the channel. For example, if the hand placement channel of a sign is changed, the entire position of the arm will have to be modified as well.
3. It may be necessary to synchronize the value of the modified channel with the other channels.
4. A component can take several successive values, if one or more of these values are changed, it is necessary to synthesize the transition between each of the values taken.

In order to select and tune our synthesis techniques, we follow an **analysis/synthesis approach** which can be broken down into several stages. First, a linguistically relevant mechanism is selected provided that some instances of this mechanism are present in the database. Those instances are extracted from the database and are considered as ground truth. The observation of this ground truth provides a relevant insight about the chosen phonological component and the behavior of the joints involved. The kinematic trajectories of those joints are studied to select and adapt the synthesis techniques that best fit the expected results. This synthesis technique is then used to reproduce the phenomenon.

The use of a captured database to extract the values taken by a phonological component (e.g., the hand configuration) and the objective of animating an avatar forces the definition of a mapping between each component (and, consequently, the corresponding annotation track) and a set of relevant joints of the avatar's skeleton. We defined a mapping (see Figure 1) that is both straightforward and has as little overlap as possible between the sets of joints corresponding to different components to have an independent control of each set of joints.

In the remainder of this paper, we aim to provide technical solutions to synthesize the phenomena listed in Table 1. The structure of the paper reflects the structure of the table with the following three sections presenting the synthesis methods developed for each phonological component to create new linguistic content.

[1] We call *channel* the set of joints corresponding to a phonological component.

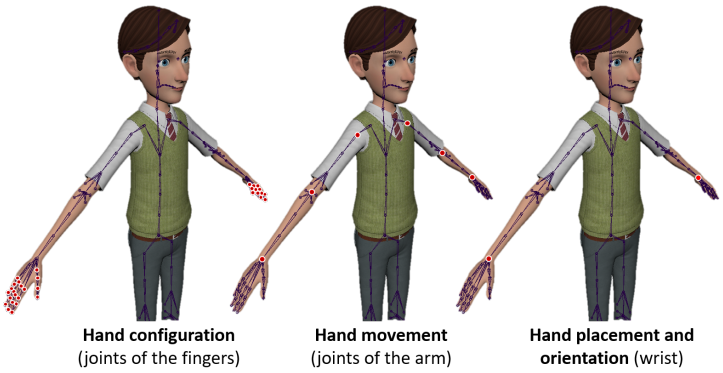

Fig. 1 Mapping between the phonological components (bold) and the avatar's joints (between parentheses).

4 Hand Placement Mechanisms

Hand Placement represents the location of the hands in the signing space. In terms of animation, it is defined as the position of the wrist joint in the signing space, which can be described by a point, a specific area, or a trajectory in the signing space. It is often specified thanks to a finite set of areas around and on the signer. Each sign has a default hand placement that corresponds to the placement of the sign in its citation form. Using the SL mechanisms of **spatialization** and **pointing gestures**, we are able to describe a wide variety of new situations by taking advantage of the possibilities of inverse kinematics combined with the retrieval of relevant signs from our database.

4.1 Spatialization: Modification of Hand Placement by Inverse Kinematics

Some signs, in their form of citation, are performed at a specific location on the body (e.g., [*PENSER*] (*to think*) on the side of the head, see Figure 2, left) or at a specific place in the signing space (e.g., [*CHIEN*] (*dog*) on the signer's side, see Figure 2, middle). The hand placement of these signs is invariant. Other signs have a default hand placement in the neutral space like [*BOL*] (bowl) (see Figure 2, right). In this case, it is possible to change the hand placement according to the context of the sign in an utterance. This mechanism of sign relocation is called **spatialization**. It is an inflection mechanism that makes it possible to precisely place objects in a scene in an absolute ("The house is on the left.") or relative way ("The house is near the swimming pool."). In the first example, the [HOUSE] will be performed on the left of the signer, while, in the second case, the same sign will be done near the place where [SWIMMING POOL] has been placed.

Spatialization therefore consists in the sole modification of the hand placement channel that we achieve thanks to inverse kinematics techniques (see Figure 3, left). To modify the hand placement and to compute the corresponding arm posture, two articulated chains were defined, one for each arm, with their respective root at the shoulder and end-effector at the wrist (see Figure 3, right).

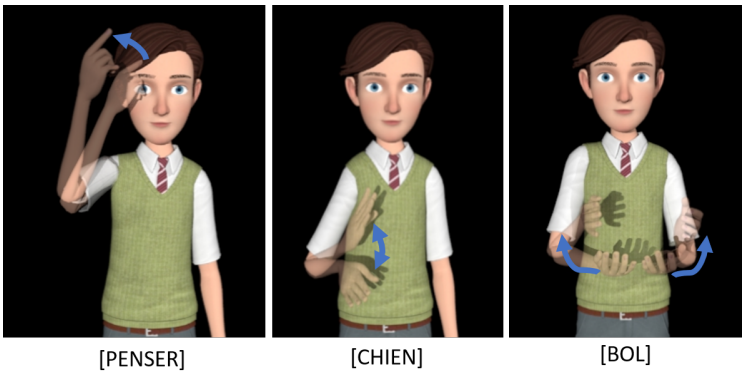

Fig. 2 Left: the sign [*PENSER*] (to think). Middle: the sign [*CHIEN*] (dog). Both signs have a precise placement in the signing space that cannot be modified. Right: the sign [BOL] (bowl) in the neutral space.

We found that traditional Jacobian-based IK methods can present too slow convergence rates and can display unstable behaviors around singularities (i.e. in situations where the end-effector cannot reach the target regardless of the changes in the angles of the articulated chain, typically in the case of out-of-reach targets). We therefore used one FABRIK solver for each chain (**?**). It provided an efficient geometrical solution and generated very few physiologically unusual poses even without any additional constraint, partially due to the fact that our chain is only composed of three joints.

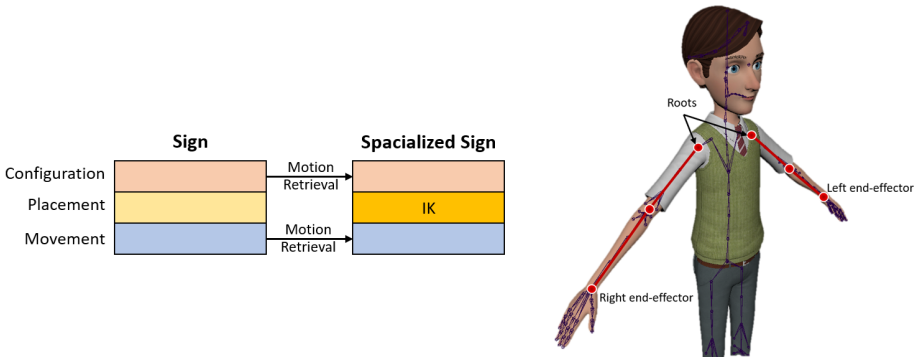

Fig. 3 Left: the spatialization principle. Right: the IK-controlled articulated chains of the two arms.

To change the hand placement channel of chosen signs while maintaining the intra-sign movement and relative placement of the two hands, we specify, at each frame, the targets of the wrists in terms of a translation between the targeted position in the world coordinates and the actual position of the wrist, still in the world coordinates, when performing the citation form of the sign. Figure 4 presents visual results of the left and upward spatializations of the sign [*BOL*] (bowl) in LSF.

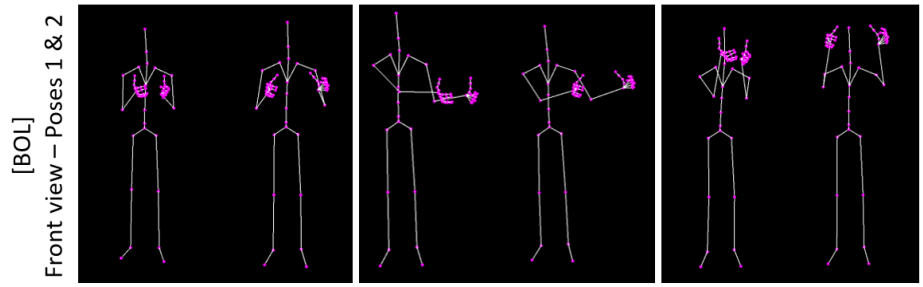

Fig. 4 The spatialization of [*BOL*] (bowl). Left: the first and last frames of the citation form of the sign [*BOL*] (bowl). Middle: a left spatialization. Right: an upward spatialization.

4.2 Pointing Gesture: Adding a Motion

Whether to designate the subject(s) of an action, to associate virtual objects to $3D$ locations in the signing space, or to refer to those locations, signers use **pointing gestures**. Index pointing gestures are the most common type of pointings and we limited our study to them as we focus on manual features but, with our phonological synthesis engine, we are capable of associating any hand configuration to pointing gestures (see Section 5.1).

In order to synthesize realistic index pointing gestures, we propose to use both inverse kinematics and interpolation techniques. We proceed as follows: the *pointing pose* for which the pointing gesture reaches its apex is computed by inverse kinematics using the FABRIK algorithm, while the *reaching* (i.e. the movement performed by the arm to reach the placement target) and *retraction* motions (i.e. the motion going back from the pointing pose) are synthesized by interpolation.

To compare the synthesized **pointing poses** with the actual ones present in our database, we first retrieved the position of the wrist at the pointing pose in the captured data, fed it to the IK solver as a target, and computed the resulting pose. We compare the real and synthesized pointing poses aiming to the same targets in Figure 5. We can see that, in the real pointing pose, the whole body is more involved than with our IK computation as we only deal with arm motions. Visually, we can see that a small difference between the actual position of the wrist and the target is tolerated by the IK model. As this difference is minor, it has no impact on the precision of the pointing gesture. Furthermore, the variability may add realism in the sense that two different initial postures will not result in exactly the same pointing pose[2].

To generate the **reaching motion**, before the pointing begins, we interpolate a neutral pose extracted from the data set and the pointing pose produced by inverse kinematics. Then, we produce the **retraction motion** by interpolating the pointing pose with another neutral pose extracted from the data set.

We tested three interpolation methods: the spherical linear interpolation (slerp) (Equation 1), the cosine interpolation (Equation 2), and the sigmoid interpolation parameterized to respect the definition domain for interpolation which is $[0; 1] \rightarrow [0; 1]$ (Equation 3). Thus, given an initial orientation q_i and a final orientation q_f

[2] It can also be interesting to add noise to achieve a similar effect.

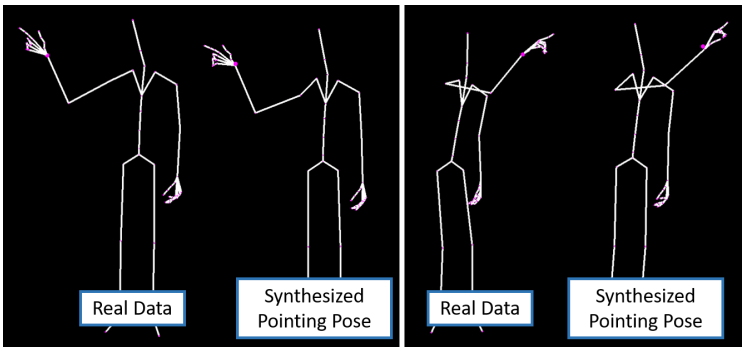

Fig. 5 Visual comparison of two real and synthesized pointing poses. The wrist targets are in purple.

of a joint j, it is possible to compute the current orientation q with respect to an interpolation weight w in the interval $[0; 1]$:

$$q_{slerp}(w) = \frac{sin((1-w)\Omega)}{sin(\Omega)} q_i + \frac{sin(w\Omega)}{sin(\Omega)} q_f \tag{1}$$

with Ω being the angle between q_i and q_f.

$$q_{cosine}(w) = (1 - \frac{1 - cos(w\pi)}{2}) q_i + \frac{1 - cos(w\pi)}{2} q_f \tag{2}$$

$$q_{paramSigmoid}(w) = (1 - \frac{1}{1 + e^{-10(w-0.5)}}) q_i + (\frac{1}{1 + e^{-10(w-0.5)}}) q_f \tag{3}$$

To compare the results of the three interpolation methods, we extracted three poses in the data set: the two poses where the hand rests near the body before (pose A) and after (pose C) the pointing, and the pointing pose (pose B). We performed the interpolation of the poses A to B and B to C to compose the full pointing gesture. Figure 6 shows the kinematic profiles for the slerp, cosine and sigmoid interpolations. The sigmoid interpolation gives the kinematic profiles closest to the ground truth. It is therefore the one we selected with the IK-generated pointing pose to create pointing motions.

A perceptual evaluation assessing the quality of synthesized animations based on hand placement mechanisms was conducted. The results are presented in Section 7.1.

5 Hand Configuration Mechanisms

The hand configuration (HC) corresponds to the shape of the hand. Concretely, at a computer animation level, it refers to the set of orientations of the finger joints. Many linguists try to establish a list of the hand configurations used in a specific sign language (for example, in LSF (**???**)). The length of this list usually varies from 30 to 80 items depending on the coarseness of the item. In either case, the hand configurations are seen as limited in numbers. It is thus possible to capture

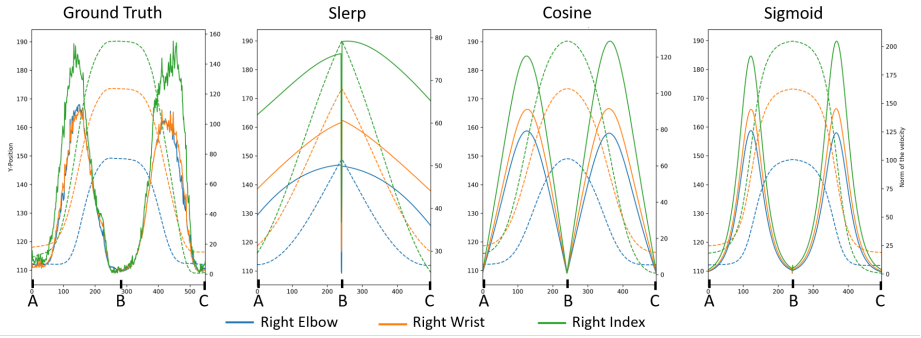

Fig. 6 The height position (dotted line) and norm of the velocity (continuous line) during one upward pointing. The real data is on the left and the different interpolated motions are shown from left to right: slerp, cosine and sigmoid interpolations. A, B and C are the timestamps corresponding to the moment when the hand rests near the body before (A) and after (C) the pointing (B).

every configuration at a low cost to have an exhaustive supply of sign language specific HC to be used in further synthesis work. The *LSF-ANIMAL* corpus (**?**) contains 48 carefully chosen hand configurations. In this section, we propose to generate new linguistically relevant content by replacing the hand configurations in specific signs (Section 5.1) and by synthesizing realistic transitions between the hand configurations (Section 5.2).

5.1 Derivative Base & Specifiers: Replacing the Hand Configurations

The HC is a relatively stable component inside a sign. Its value rarely varies during the production of individual signs and can generally be labeled unambiguously following pre-established categories. New SL can thus be synthesized by modifying the values taken by the HC inside isolated signs. The signs that can be created this way are limited by the initial database and by the language itself: it is necessary to have an *a priori* knowledge of the language as not all the combinations sign/configuration exist.

Three LSF mechanisms are particularly suited to such transformation: some specific **derivative bases**, **shape specifiers**, and **proforms**. It is assumed, in those three cases, that the nature of the HC is discriminative and that new signs can be created by modifying it: given an initial sign in the database, it is possible to create new signs by retrieving a different HC in the database and, then, by overwriting the initial HC with the new one in the sign. In this paper, we only treated the derivative bases and shape specifiers. The synthesis of proforms, which consists in doing specific HC to impersonate the subject or object of an action (a flat hand represents a car, a raised thumb a cyclist, etc.), constitutes perspectives for future work.

A **derivative base** designates a set of signs with the same value for one phonological component which has a unity of meaning. In the case of HC replacement, we only target the derivative base with the same motion and placement. In *LSF-ANIMAL*, we have such derivative bases: the signs [*ESCARGOT*] (snail) and [*LIMACE*] (slug) are part of the same derivative base grouping slow rampant

animals. The signs in this derivative base have the same slow yawing motion and the dominant hand is placed slightly above the non-dominant hand. They only differ by the hand configurations: configuration of the 'H' or 'Y' for [*ESCARGOT*] (snail) and 'U' for [*LIMACE*] (slug). For our experiments, we replaced the 'Y' hand configuration in [*ESCARGOT*] (snail) to create an 'H' configuration snail and a 'U' configuration slug. The 'H' and 'U' configurations originate from the isolated HC sequences of the *LSF-ANIMAL* corpus (**?**). Concretely, we overwrite the 'Y' configuration of the skeleton during the realization of the snail sign with the orientation of the finger joints for the 'H' and 'U' configurations. The results are visible on Figure 7.

Fig. 7 Replacement of the hand configuration in the sign [*ESCARGOT*] (snail). Left: original 'Y' configuration. Middle: replacement with an 'H' configuration to create a different [*ESCARGOT*] (snail). Right: replacement with a 'U' configuration to create the sign [*LIMACE*] (slug).

In **shape specifiers**, the HC will vary depending on the object of the action. In the *LSF-ANIMAL* corpus, we have an example of shape specifiers in the gait of our different animals. A character's gait can be precisely described by using the hands and arms of the signer to represent the legs of the described thing. Gaits are both the results of shape specifiers and role shift. Indeed, different animal gaits can be generated by changing the hand configurations (e.g., a cat's walk can be changed to a lion's walk by changing the 'U' to a '5_folded' configuration): here, the hand configuration is the shape specifier. However, in the case of gaits, hand configuration is not the only thing that varies; the style of the arm movement is decisive when describing a character's type of walk: this is the expression of the role shift part that is not treated in this paper. We tested HC replacement for three other gaits: we replaced the 'U' configuration of the cat's gait by a '3' configuration to obtain a chicken's gait, a 'S' configuration to obtain a cow's gait and a '5_folded' configuration to obtain the lion's walk (see results on Figure 8).

5.2 Dactylology: Linking the Configurations

Dactylology, also called fingerspelling, is the process of spelling a word by using a dactylological alphabet. The French one consists of 19 hand configurations that – when held in certain orientations or are produced with certain movements – represent the 26 letters of the French alphabet. In LSF, spelling is mainly performed with the dominant hand in the neutral space: the placement component is thus fixed, modifications of orientation and movements are minimal.

A fingerspelled word is not a sign in itself. However, some signs are **derived from fingerspelled words** like [*OK*] (hand configuration of the 'O' followed by the 'K'), [*WEEK-END*] ('W' followed by 'E' with a rectilinear movement) in LSF,

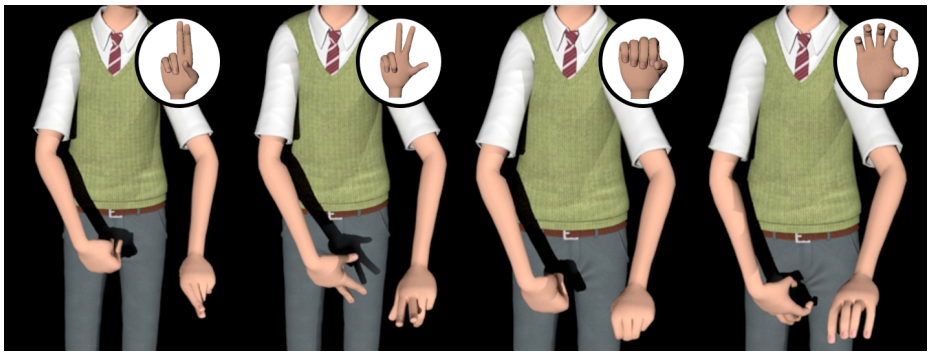

Fig. 8 Replacement of HC in gaits: the original motion is the cat's walk (left) with the 'U' configuration, the three others are synthesized (from left to right: chicken, cow and lion's walk).

or the sign [*LSF*] itself ('L', 'S', 'F' configurations with a descending and then ascending rectilinear hand movement, see Figure 9).

While Section 5.1 only deals with static hand configurations, fingerspelling and signs derived from dactylology require a finer control of the transformations of the hand configuration over time. For such signs, it is necessary to manage the profile and the timing of the transition from one configuration to another in a way that is both correct and realistic.

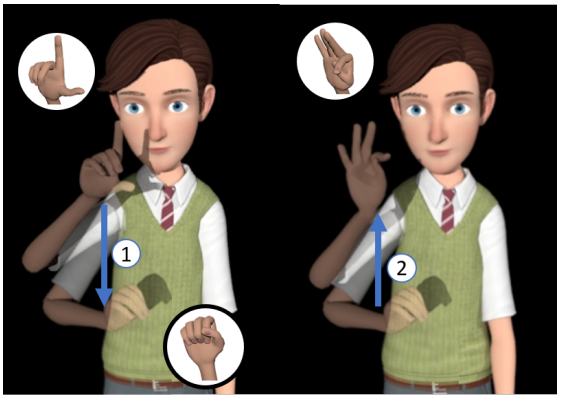

Fig. 9 The sign [*LSF*]: 'L', 'S' and 'F' configurations with a descending (1) and then ascending (2) hand movement.

To link two HC, we therefore propose to use static hand configurations present in the isolated HC sequences of the *LSF-ANIMAL* corpus (**?**) and to synthesize the transition procedurally. It is thus a matter of synthesizing the variation of the hand configuration from a fixed 1-frame initial configuration to a 1-frame final configuration with optionally 1-frame intermediate configurations.

The three types of interpolation presented in Section 4.2 were applied on examples of signs derived from dactylology. Figure 10 presents the norm of the position and the velocity of the tip of the fingers for the dactylological sign [*DODO*] (to nap) (hand configuration of the 'D' followed by the 'O' and repeated once) for the ground truth and the different interpolations. Whether for the velocity or the po-

sition of the fingers over time, the sigmoid interpolation seems to give the results most similar to the ground truth.

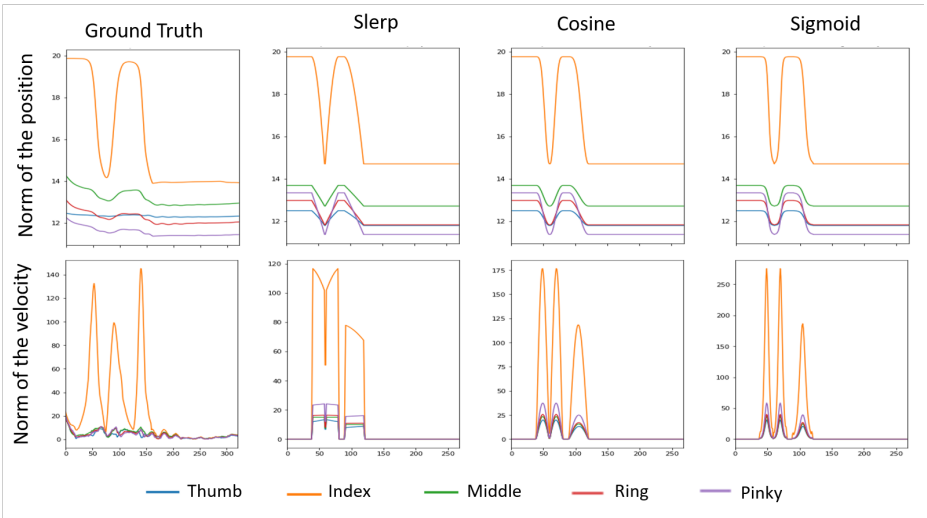

Fig. 10 The position (first row) and norm of the velocity (second row) of the tip of the fingers in the dactylological sign [*DODO*] (to nap). The columns represent, from left to right, the curves of the ground truth, slerp, cosine and sigmoid interpolations.

A perceptual evaluation assessing the quality of synthesized animations based on hand configuration mechanisms was conducted. The results are presented in Section 7.2.

6 Hand Movement Mechanisms

Hand movement corresponds to the overall trajectory of the wrist during a sign and, sometimes, to secondary movements (i.e. small movements of the fingers) also performed during a sign. In this section, we only study the former as we consider the latter to be closer to a transformation of the hand configuration than to a movement phenomenon. Hand movement can have different roles depending on the linguistic level considered. At a phonological level, hand movement is an *articulatory motion* that can be modified through different operators.

We define the articulatory motion as the hand movement performed as part of an isolated sign (unlike inflection mechanisms involving the hand movement, like *motion paths* in the description of situation using proforms, or *iconic flexions* to precisely describe the motion of an object). By modifying the articulatory motion of the hands, it is therefore possible to create new signs absent from the original database. To that end, we propose to edit the hand movements when they act as articulatory motions by treating them as time series. Indeed, time plays a predominant role in the production of SL content, as a gesture can be seen as a series of spatiotemporal targets to be reached. If the temporality is changed, the sign will be modified: it will lose its meaning or gain in subtlety. The execution of these mechanisms for the animation of an avatar involves motion editing processes. A

sign will take on a different meaning or style through temporal inversion, mirroring or repetition of the hand movement.

Temporal Inversion: In LSF, some signs have an opposite that can be achieved by temporally reversing the movement. For example, the sign for [*CLAIR*] (light) can be changed to [*FONCÉ*] (dark) if it is played backwards (see visual results on Figure 11). The same goes for [*AIMER*]/[*NE PAS AIMER*] (like/don't like) or [*DONNER*]/[*PRENDRE*] (give/take) (**?**). The presence of signs with potential opposite by inversion in the initial base is very interesting for the enrichment of the base.

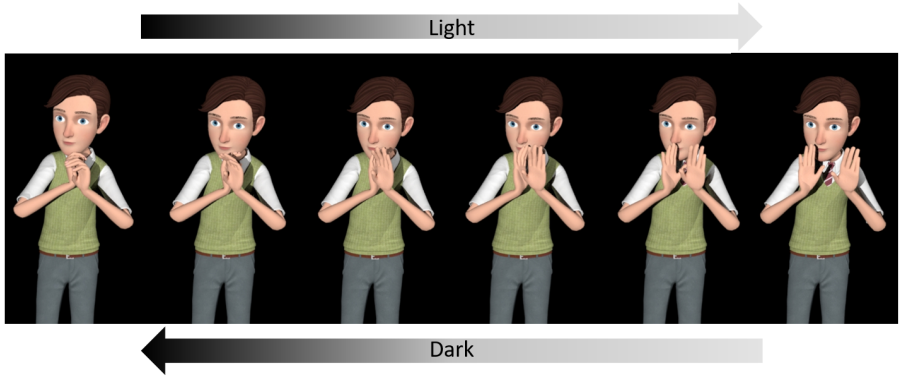

Fig. 11 Synthesis of the sign [*FONCÉ*] (dark) with the temporal inversion of [*CLAIR*] (light).

Swap and Mirror: In order to increase the number of variations in the database, it is also possible to transform the signs performed with the right hand (resp. the left hand) into left-handed signs (resp. right-handed signs). A *hand swap* operator makes it possible to exchange the hand movement performed by the dominant and non-dominant hands. This *swap* operator is also useful for utterance synthesis and, particularly, for the parallel production of two signs, as signs can be done with either hand, in particular if the preferred hand is doing another sign. In addition, some signs can be done with either one or two hands (e.g., the signs [*WEEK-END*] or [*FIN*] (end) in LSF). Generating the two-handed version from the single-handed one can be done with the *mirror* operator by copying the movement of one arm on the other arm. The swapped and mirrored signs thus generated keep the realism of the captured motions. Examples of application of the hand swap and mirror operators are shown on Figure 12.

Repetition: A repetition in movement can have several meanings: it can describe the repetition of an action or an event, or add the notion of weariness. The sentence "I work all the time, I'm tired of it" can potentially be executed with only the sign [*TRAVAIL*] (work) repeated many times. The amplitude and number of repetitions will change according to the desired meaning[3]. Repetition is also a means of constructing new signs. For instance, the sign [*PUNIR*] (to punish) repeated twice results in the sign [*TRAVAIL*] (work) (see Figure 13). In order to have a smooth transition between two repetitions of the same motion, a short slerp interpolation is added between the repeated instances.

[3] Not to mention the overall attitude and facial expression that are not part of this work.

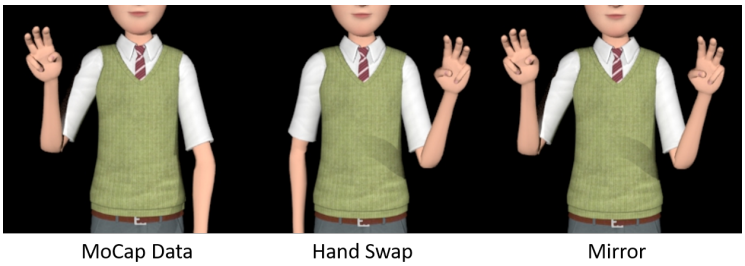

Fig. 12 Results of the application of the swap (middle) and mirror (right) operators for the sign [*WEEK-END*]. The original version of the sign is on the left.

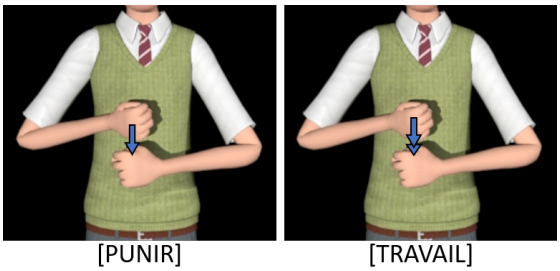

Fig. 13 The sign [*PUNIR*] (left) repeated twice gives the sign [*TRAVAIL*] (right).

The swap, mirror, temporal inversion and repetition operators presented in this section can be combined to create new content (e.g., the sign [*ÊTRE DÉGOUTÉ*] (to be disgusted) can be built by temporally inverting and, then, repeating the sign [*AIMER*] (to like)).

7 Perceptual Evaluation

In order to assess the quality of our synthesis, we perceptually evaluated a subset of the synthesized motions. We made two perceptual evaluations: (i) a first one to evaluate the quality of our synthesized spatialization instances and pointing gestures (hand placement mechanisms) and (ii) a second one to assess the quality of the synthesized derivative base, shape specifiers and dactylology mechanisms (hand configuration mechanisms). Both evaluations were performed using online anonymous questionnaires with questions asked in both written French and LSF.

7.1 Evaluation of the Hand Placement Mechanisms

In the first perceptual evaluation, we aimed to judge the quality of our synthesized spatialization mechanisms and pointing gestures. To this end, we formulated the three following hypotheses:

H_1: The placement of synthesized spatialized instances is as **recognizable** as the placement of ground truth spatialized instances.

H_2: Synthesized spatialized instances are as **realistic** as ground truth instances.
H_3: Pointing gestures performed using the sigmoid interpolation are more **realistic** than pointing gestures performed using linear interpolation.

7.1.1 Design of the Evaluation

We designed an online questionnaire in order to test those three hypotheses. The questionnaire was divided into two parts: the first one for the evaluation of the spatialized mechanisms (hypotheses H_1 and H_2) and the second one for the evaluation of the realism of the pointing gestures (H_3).

In order to assess the quality of the synthesized spatialization mechanisms, we built 10 videos showing an avatar describing, with the spatialization of the sign [*BOL*] (bowl), different placements of a bowl in a scene. 5 of those 10 videos were done as a simple play-back of captured data and were considered as ground truth. The other 5 videos showed the same spatialization instances synthesized with our technique using inverse kinematics as described in Section 4.1. For each video, we asked the participants to select the described placement of the bowl among 6 possibilities (hypothesis H_1, see Figure 14) and to rate the realism of the movements on a 5-point Likert scale (hypothesis H_2). We used one of the 5 ground truth videos as a training example to explain the task to the participants. Ultimately, the participants evaluated 8 spatialization videos: 3 ground truth videos[4] and 5 videos showing synthesized results.

Fig. 14 One of the task in the perceptual evaluation of the spatialization technique. A video is shown to the participants (left) who must select the placement of the spatialized instance among 6 possibilities (right).

To compare the results of the linear and sigmoid interpolations for pointing gestures, we built 9 videos, each showing an avatar performing a sequence of 2 to 5 different pointing gestures. Like previously, 3 of those 9 videos were a play-back of the captured data and were considered as ground truth. The 6 remaining videos were synthesized using IK with the same targets as the ground truth pointing gestures with linear interpolation (3 videos) or sigmoid interpolation (3 videos), both for reaching and retraction motions. For each video, we asked the participants to rate the realism of the pointing gesture on a 5-point Likert scale (hypothesis H_3).

[4] One of the 4 remaining ground truth videos was removed from the questionnaire beforehand and was not showed to the participants as it contained an artefact.

7.1.2 Results

For this first evaluation, we collected the results from 75 participants, 27 men and 48 women with an average age of 40.8 years old (+/- 14, 4 years). Among the participants, 48 were born deaf, 10 had become deaf during their lifetime, 6 were hearing-impaired and 11 were hearing people. In addition, the participants were asked to assess their level of French Sign Language (*no knowledge of LSF*: 1 participant, *beginner*: 6, *quite good*: 11, *good*: 15, *very good*: 16, *native*: 24, or *interpreter*: 2). We assumed that the 57 participants with a *good* level of LSF and above (*very good*, *native* and *interpreter*) were most sensitive to the subtle variations between the different ways of synthesizing the movements. We therefore only considered their answers in the remainder of this section.

To evaluate the **precision of our spatialization technique**, we used the results of the recognition task: in each video, the sign [BOWL] was performed in a different location and the participants had to select the placement of the sign among 6 possibilities. Surprisingly, the recognition rate of the placement of the synthesized sequences was much higher (85.96%) than the recognition rate of the play-back sequences (63.15%). We explain that for two main reasons :

1. There were more sequences of synthesized instances (5) than of play-back sequences (3). As such, a mistake in the play-back sequences had a higher cost than for the synthesized instances.
2. The play-back sequences presented more variations in the shape of the bowl than the synthesized instances that reproduced the citation form of the sign at different places. Those differences in the way of performing the sign itself may have had an impact on the perception of the placement of the bowl for the participants.

Still, with a recognition rate of 85.96% for the synthesis against 63.15% for the ground truth, we found that the placement of the synthesized instances using our technique for spatialization was at least as good as the ground truth. We therefore validated the H_1 hypothesis stating that "The placement of synthesized spatialized instances is as recognizable as the placement of ground truth spatialized instances.".

To evaluate the **realism of the synthesized spatialization sequences**, we asked the participants to rate the realism of each of the 8 videos on a 5-point Likert scale. Still considering only the 57 participants with a *good* level of LSF and above, we obtained a mean rating of 3.614/5 (std of 1.262) for the play-back sequences and a slightly higher mean value of 3.839/5 (std of 1.177) for the synthesized sequences (see Figure 15, left). We found no significant difference between the realism ratings of the ground truth and synthesized sequences (*p-value* of 0.031 with the Mann-Whitney test)[5]. We therefore validated the H_2 hypothesis stating that synthesized spatialized instances are as realistic as the equivalent ground truth instances.

To test the **realism of the pointing gestures**, we asked the participants to rate each of the 9 videos on a 5-point Likert scale. Considering only the ratings of

[5] We considered a significant difference for a *p-value* < 0.01.

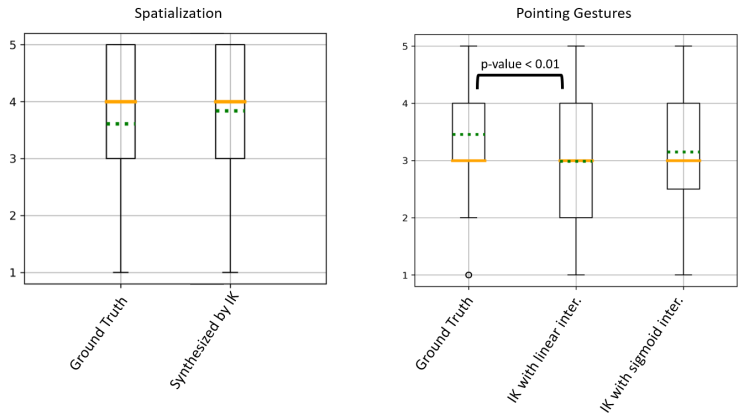

Fig. 15 Left: Realism rating of the ground truth and synthesized spatialization instances (left) and ground truth and synthesized pointing gestures using IK and the linear or sigmoid interpolation (right). The median is the orange line, the mean is the dotted green line, the whiskers go to 1.5 multiplied by the interquartile range. We only considered the answers of the participants with a *good* level of LSF and above.

the participants with a *good* level of LSF and above, we found that the mean rating for the ground truth pointing gestures was 3.456/5 with a standard deviation of 1.14. The linear interpolation pointing gestures obtained a mean rating of 2.988/5 (std : 1.27) and the mean rating for the sigmoid interpolation pointing gestures was 3.152/5 (std : 1.155) (see Figure 15, right). Using the Kruskal-Wallis test for non-parametric data and the pairwise Mann-Whitney statistical tests, we found that there was a significant difference between the realism ratings of the ground truth sequences and the ones of the linear interpolation synthesis sequences (*p-value* of $2.48e{-}4$). However, we found no significant difference between the realism ratings of the ground truth sequences and the ones of the sigmoid interpolation synthesis sequences (*p-value* of 0.081). This perceptual evaluation therefore showed that pointing gestures synthesized following our technique combining IK and sigmoid interpolation gives realistic results. It also showed that the type of interpolation used in the synthesis is relevant and should be considered carefully, as a simple linear interpolation was seen as significantly less realistic than the sigmoid interpolation. Those results confirmed the H_3 hypothesis stating that "Pointing gestures performed using the sigmoid interpolation are more realistic than pointing gestures performed using linear interpolation.".

7.2 Evaluation of the Hand Configuration Mechanisms

In order to assess the accuracy and realism of the synthesis of hand configuration mechanisms, we performed a second perceptual evaluation. We formulated the four following hypotheses :

H_4: The synthesis of signs by replacing the HC provides results as **realistic** as play-back data.

H_5: The synthesis of signs derived from dactylology provides results as **accurate** as play-back data.
H_6: The synthesis of signs derived from dactylology provides results as **realistic** as play-back data.
H_7: There is a significant difference between the use of the slerp and the sigmoid interpolation in the quality (**accuracy** and/or **realism**) for the synthesis of signs derived from dactylology.

7.2.1 Design of the Evaluation

To answer those four hypotheses, we designed an experiment showing 20 videos put in a random order. Those 20 videos are described below and in Table 2.

First, to evaluate our synthesis method based on the replacement of the hand configuration, we created two ground truth motion sequences: one containing a play-back of the captured data for [*ESCARGOT*] (snail) performed with a 'Y' hand configuration and another with a play-back of the captured data for the cat's walk. From the "snail" video, we synthesized two new sequences by replacing the 'Y' hand configuration: another sign [*ESCARGOT*] (snail) with an 'H' hand configuration and a sign [*LIMACE*] (slug) with a 'U' hand configuration. In addition, we constructed the "rooster's walk" by replacing the 'U' hand configuration of the "cat's walk" with a '3' configuration as detailed in Section 5.1. We thus obtained 5 videos.

Then, to assess the quality of the interpolation method to link the configurations, we chose 5 signs derived from dactylology: [*LSF*], [*WEEK-END*], [*OR*] (gold), [*OK*] and [*SALON*] (living-room). We created 15 videos showing the avatar performing three different versions of each of those 5 signs: the first version is a simple play-back of the captured data and was considered as ground truth, the second version was created using our method with a slerp interpolation and the third version was created using our method with a sigmoid interpolation.

For each video, participants were asked to select the gloss corresponding to the sign that was performed among 24 visually close possibilities (hypotheses H_5 and H_7) and to rate the realism of the movements on a 5-point Likert scale (hypotheses H_4, H_6 and H_7). The participants could watch the videos as many times as they wished.

#	Sign	Versions (number of videos)
1-2	[*ESCARGOT*]	ground truth and synthesized (2)
3	[*LIMACE*]	synthesized (1)
4	Cat's walk	ground truth (1)
5	Rooster's walk	synthesized from the cat's gait (1)
6-8	[*LSF*]	Ground truth, slerp and sigmoid (3)
9-11	[*SALON*]	Ground truth, slerp and sigmoid (3)
12-14	[*WEEK-END*]	Ground truth, slerp and sigmoid (3)
15-17	[*OR*]	Ground truth, slerp and sigmoid (3)
18-20	[*OK*]	Ground truth, slerp and sigmoid (3)

Table 2 Description of the content of the 20 videos of the perceptual evaluation of the hand configuration mechanisms.

7.2.2 Results

For this second evaluation, we collected the results from 53 participants, 22 men and 31 women with an average age of 37.2 years old (+/- 13, 4 years). Among the participants, 32 were born deaf, 7 had become deaf during their lifetime, 1 was hearing-impaired and 13 were hearing people. As for their level of French Sign Language: 8 were *beginners*, 6 were *quite good*, 8 were *good*, 12 *very good*, 18 had a *native* level, and 1 was an *interpreter*. Like for the evaluation of the hand placement mechanisms, we chose to only consider the participants with a *good*, *very good*, *native* and *interpreter* level of LSF which amounted to 39 participants.

To evaluate the **realism of the hand configuration replacement techniques** of Section 5.1, we considered the ratings given on a 5-point Likert scale of the 5 videos corresponding to the gaits and snail/slug sequences. Table 3 shows the mean realism ratings per sequence. We can notice that the rooster's gait, done by replacing the hand configuration of the cat's gait, obtained slightly better ratings than its play-back counterpart. However, we did not detect any significant difference between the ratings of the two sequences (*p-value* of 0.1988). In a similar way, the synthesized snail obtained better ratings than the ground truth snail. Using the pairwise Mann-Whitney tests on the ground truth snail, synthesized snail and synthesized slug, we found a significant difference between the synthesized slug and synthesized snail (*p-value* of $2.12e^{-3}$). We think that this difference comes from the fact that the slug sign that we generated is not common and deaf people may prefer other variants. In a similar way, the 'Y' hand configuration ground truth snail, as performed by our signer, was less known than the H configuration synthesized snail (see Figure 7). This could have had a negative impact on the realism ratings of the synthesized slug and ground truth snail. No significant difference between the ground truth and results of the syntheses were found. Even if there are difference in quality in the synthesized instances, they are deemed to be as realistic as the ground truth. We thus validated the H_4 hypothesis ("The synthesis of signs by replacing the HC provides results as realistic as play-back data.").

Sequence	Mean rating (/5)	Std
Cat's gait (GT)	2.974	1.25
Rooster's gait (synth)	3.231	1.143
[SNAIL] (GT)	3.103	1.215
[SNAIL] (synth)	3.538	1.195
[SLUG] (synth)	2.718	1.28

Table 3 Mean and standard deviation of the realism ratings for the ground truth (GT) and synthesized (synth) sequences.

To evaluate the **accuracy of the synthesis of signs derived from dactylology** (hypotheses H_5 and H_7), we used the 15 remaining videos and the results of the recognition task: a sign derived from dactylology was performed in each video (either a play-back of the captured data, or synthesis results using either the slerp or the sigmoid interpolation) and the participants had to select the meaning of the sign among 24 close possibilities. Figure 16 shows the recognition rate of each sign with respect to the type of sequence ("Ground Truth" (play-back sequences),

slerp interpolation or sigmoid interpolation). The recognition rate of the sign [*OR*] (gold) is slightly higher using the slerp interpolation while the recognition rate of the sign [*OK*] or [*WEEK-END*] is slightly higher using the sigmoid interpolation. Using the Kruskal-Wallis test, we found no significant difference between the ground truth, slerp and sigmoid interpolations. The absence of difference between the ground truth and the two synthesis techniques makes it possible to validate the H_5 hypothesis : "The synthesis of signs derived from dactylology provides results as accurate as the ground truth.". However, as there was no significant difference between the slerp and sigmoid techniques in terms of recognition rate, we could not validate the H_7 hypothesis.

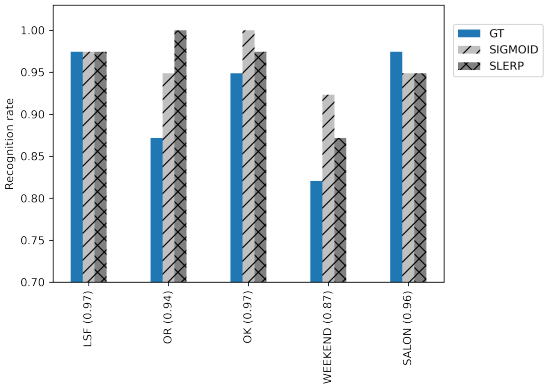

Fig. 16 Recognition rate of each sign with respect to the type of sequence: ground truth, slerp interpolation or sigmoid interpolation. The number between parenthesis is the mean recognition rate for each sign (all techniques considered).

Finally, to evaluate the **realism of the synthesized signs derived from dactylology**, we relied on the realism ratings given on a 5-point Likert scale by the participants for each video. The left part of Figure 17 shows the box plots of the realism ratings per sign and per technique while the signs are grouped in the right part of Figure 17 to give a more general view of the results of the three types of sequences. We observed that the ground truth (mean value of 3.026/5, std of 1.279), slerp (mean value of 3.082/5, std of 1.217) and sigmoid (mean value of 3.067/5, std of 1.325) sequences did not present any significant difference. This observation allows us to validate the H_6 hypothesis ("The synthesis of signs derived from dactylology provides results as realistic as the ground truth") and to reject the H_7 ("There is a significant difference between the use of the slerp and the sigmoid interpolation").

Indeed, the participants did not perceive any difference between the use of the slerp and of the sigmoid interpolation. Only affecting the way the fingers are deployed, the impact of this interpolation seems limited for dactylology. However, even if the difference between the slerp and sigmoid interpolation is subtle on the avatar animations presented during this evaluation, the kinematic profiles of the movements created with the sigmoid are visually much closer to the ground truth (see Figure 10). We thus think that on longer sequences and for eyes accustomed to avatars, the sigmoid should be better appreciated than a simple slerp. Moreover,

the impact of this interpolation is only considered here in the context of dactylology where rapid transitions between manual configurations are performed. However, it might be interesting to extend the analysis of the opening or closing of the hand to other signs and utterances, particularly for the study of expressive gestures or prosody in SL.

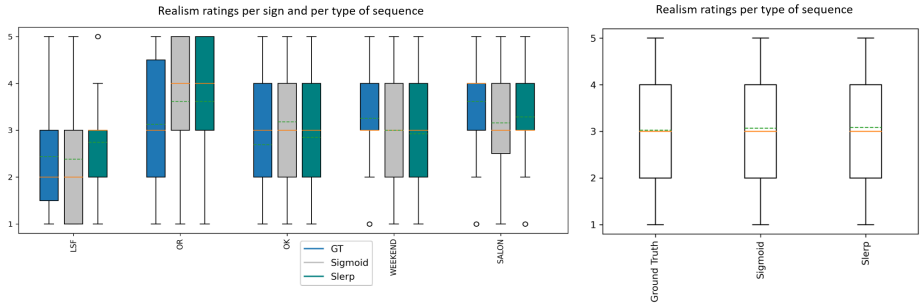

Fig. 17 Left: realism ratings of the ground truth (in blue) and synthesized signs using the slerp (green) and sigmoid (gray) interpolation with respect to each sign. Right: Realism ratings grouped by type of sequences (ground truth, slerp or sigmoid interpolation). The median is the orange line, the mean is the dotted green line, the whiskers go to 1.5 multiplied by the interquartile range.

8 Conclusion and Perspectives

A large part of the existing work in the field of signing avatar animation mainly uses, for sign synthesis, either the playback of real motion data or pure procedural synthesis. In the first case, the movements of the avatar keep the realism of the real movements but the number of signs that can be generated is limited by the initial database, and the transformation of signs requires complex signal processing techniques. In the second case, the variety of signs that can be synthesized is far less restricted but the construction of new signs can require a time-consuming process, and the resulting signs are robotic and unrealistic.

We have proposed a hybrid system that takes advantage of both approaches in order to create new signs, absent from an annotated database of initially captured movements, but retaining the realism properties of real movements. For this, we rely on a vision centered around phonological components as defended by linguistic work and at the base of many sign representation systems like *HamNoSys* (**?**). We have described different techniques to build new signs in their citation form as well as inflection phenomena by modifying the values taken by the individual phonological components. In addition to simple recombination, we have implemented pure synthesis techniques, namely inverse kinematics and interpolation techniques that we have adapted to be as close as possible to the resulting ground truth. A video showing our synthesis results is available at `http://sltat.cs.depaul.edu/2019/naert.mp4`.

We performed two qualitative evaluations of our work to assess the accuracy and realism of the proposed techniques for the generation of animations involving hand placement and hand configuration mechanisms. We collected the answers of 75 participants for the first evaluation and 53 participants for the second one, a

majority of whom were born deaf. In those evaluations, we compared our results with play-back data that we considered as ground truth. We showed that the results of the proposed synthesis techniques were not significantly different from those of the ground truth, which means that our synthesized animations cannot be differenciated from the ground truth and may be substituted to the *MoCap* data to enrich the original motion database. The evaluation of the hand movement synthesis techniques will be the focus of future work.

In the context of this paper, we focused on three phonological components: hand configuration, hand placement and hand movement. However, hand orientation and other non-manual components of SL are crucial for the proper understanding of signs. The analysis of their role and synchronization with arm movements is a promising perspective for the continuation of this work.

We also intend to extend our synthesis system by implementing new, higher-level spatiotemporal mechanisms that would lead to the specification of indicating verbs, proforms and size specifiers and, more generally, highly iconic signs and utterances.