The SignCom System for Data-Driven Animation of Interactive Virtual Signers : Methodology and Evaluation

Sylvie Gibet, Nicolas Courty, Kyle Duarte, Thibaut Le Naour

The *SignCom* System for Data-Driven Animation of Interactive Virtual Signers: Methodology and Evaluation

SYLVIE GIBET, NICOLAS COURTY, KYLE DUARTE, and THIBAUT LE NAOUR
Université de Bretagne-Sud, Laboratoire VALORIA

In this paper we present a multichannel animation system for producing utterances signed in French Sign Language (LSF) by a virtual character. The main challenges of such a system are simultaneously capturing data for the entire body, including the movements of the torso, hands, and face, and developing a data-driven animation engine that takes into account the expressive characteristics of signed languages. Our approach consists of decomposing motion along different channels, representing the body parts that correspond to the linguistic components of signed languages. We show the ability of this animation system to create novel utterances in LSF, and present an evaluation by target users which highlights the importance of the respective body parts in the production of signs. We validate our framework by testing the believability and intelligibility of our virtual signer.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language generation*; I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism—*Animation*; J.5 [**Arts and Humanities**]: —*Linguistics*

General Terms: Algorithms,Design,Experimentation,Human Factors,Languages

Additional Key Words and Phrases: Communicative gestures, data-driven animation, multichannel animation, multimedia generation, multimodal corpora, signed language gestures

1. INTRODUCTION

For some time now, the computer animation and signed language linguistics communities have been jointly interested in developing signing avatars capable of realistic communication in signed languages; the *SignCom* project is one of many in this category of research.

As with all signing avatar projects, *SignCom* focuses on the nonverbal modalities of human-machine interaction, particularly human-humanoid interaction. More specifically to *SignCom*, the project seeks to build interaction between users and virtual agents communicating in French Sign Language (LSF), and thus engaging in real-time dialog. This is achieved by the human user signing towards a camera by which the system recognizes his/her signs, and by the virtual agent providing culturally- and linguistically-acceptable responses in behavior and sign, respectively. In this paper, we will present the sign generation part of this interactive system, with the specific goal of producing real-time, novel, and realistic LSF utterances.

Most signing avatar projects adopt synthetic animation techniques for their virtual agents, methods that have not as yet been able to convince audiences with their overt realism. In our work however, we draw on data-driven animation techniques, i.e., those that record human motion, such as motion capture. We believe that data-driven virtual signers will have more fluid and thus more convincing signing styles than their synthetic counterparts. Just the same, animating expressive

virtual signers in an interactive context becomes tedious, mostly for these reasons:

i the linguistic (read phonological) structure of signs is still widely debated in the sign language linguistic community, thus modeling particular aspects of signed languages using such elements may sometimes fail in forming new signs or utterances;

ii animation methods are made more complex by the multi-channel nature of gestures found in signed languages and by the need for real-time output imposed by the interactive nature of the application; and

iii the intended user group (i.e., Deaf[1] people) are known to be critical of misuse of their language, requiring thorough evaluation of the system's output.

The major contributions of this paper are in the direction of the last two points, i.e., the animation methods we use for the virtual signer, and our evaluation of the system through a survey of our target audience. In all, we present an original data-driven animation system dedicated to linguistic interaction between humans and virtual agents using French Sign Language. This system uses both accepted components from the animation community as well as original modules, such as our streaming architecture for motion retrieval or our regression technique for facial animation. The system is preliminarily evaluated to quantify the system's acceptability and the pertinence of our technical choices.

The paper is organized as follows. We begin by describing the historical and technical context for creating a signing avatar, including a discussion on procedural versus data-driven models, and give an overview of the *SignCom* project in Section 2. Section 3 will describe our data capture sessions, and discuss the design and annotation of our signed data following a current linguistic theories and practices; this data is stored in databases of motion and semantic content for later retrieval. Then, in Section 4, we will detail our animation modules, specifically those that handle simultaneous corporal animation, facial expressions, and gaze direction, and discuss how these modules make use of the data stored in the previous section. Ultimately, we ask target users to evaluate the ability of our generation system to produce novel utterances in French Sign Language, isolating motions by channel within various scenarios, and comparing synthesized sequences with playback sequences; results are outlined in Section 5 and discussed in Section 6.

2. CONTEXT AND MOTIVATION

As Deaf people generally do not have access to the sounds of spoken languages, signed languages are their native languages, being fully accessible through the visual modality. Also, by nature of living in a hearing world, Deaf people are necessarily bilingual, reading signs and gaining world knowledge by reading the written language of their hearing compatriots. However, spoken/written language fluency varies considerably among members of any single Deaf community, and relying solely on written text or subtitles can be challenging.

[1]We follow orthographic convention in this paper, using a lowercase *deaf* to describe the physical condition of deafness, and an capitalized *Deaf* to refer to the linguistic and cultural traditions of a people group.

Thus, novel interactive systems for communication in signed languages have been developed in recent years to improve Deaf accessibility to various media. The most current and linguistically-native assistive technologies center around the generation of signed languages using virtual humans, or avatars. Importantly, avatars present an exciting opportunity to model interactions based on the desires of the interlocutor, to provide on-the-fly access to otherwise inaccessible content, and even to anonymize interactions between interlocutors.

We review here some of the technologies used to animate virtual communicative agents, separating descriptive and procedural methods from data-driven animation methods, then we present the main technologies used for virtual signers, and finally we describe the main objectives of the *SignCom* project, which incorporates a fully data-driven approach for animating a virtual signer.

2.1 Descriptive and Procedural Methods for Embodied Communicative Agents

Gesture taxonomies have been proposed early on in [McNeill 1992] and [Kendon 1993], some of which require the identification of specific phases in co-verbal gestures and signed language signs [Kita et al. 1997]. Recent studies on expressive gesture rely on the segmentation and annotation of gestures to describe the spatial structure of a gesture sequence, or transcribe and model gestures with the goal of later re-synthesis [Kipp et al. 2007].

A growing number of systems provide for the animation of embodied conversational agents (ECAs). In such systems, crossing linguistics, artificial intelligence, and psychology with computer animation, gestures have been described at behavioral-planning levels and generated with animation engines.

Regarding high-level gesture specification, historical and current methods range from formalized scripts to dedicated gestural languages. The Behavior Expression Animation Toolkit (BEAT), as one of the first systems to describe the desired behaviors of virtual agents, uses textual input to combine gesture features for generation and synchronization with speech [Cassell et al. 2000]. XML-based description languages have been developed to describe various multimodal behaviors, some of which are dedicated to complex gesture specification [Kranstedt et al. 2002], describe style variations in gesture and speech [Noot and Ruttkay 2005], or introduce a set of parameters to categorize expressive gestures [Hartmann et al. 2006]. Vilhalmsson et al. and Kopp et al. have defined the BML unified framework containing several levels of abstraction, which interprets a planned multimodal behavior into a performed behavior, and may integrate different planning schemas and controllers [Vilhalmsson et al. 2007; Kopp et al. 2006]. More recently, the real-time system EMBR introduces a new layer of control between the behavioral level and the procedural animation level, thus providing the animator a more flexible and accurate interface for synthesizing nonverbal behaviors [Héloir and Kipp 2010].

Passing from the specification of gestures to their generation has given rise to a few publications. Largely, the authors of these methods desire to translate a gestural description, expressed in any of the above-mentioned formalisms, into a sequence of gestural commands that can be directly interpreted by a real-time animation engine. Most of these concern pure synthesis methods, for instance by using inverse kinematics techniques, such as in [Tolani et al. 2000; Kopp and Wachsmuth 2004].

These approaches using high-level specification languages coupled with procedural animation methods allow for building consistent and precise behaviors, and can account for specific constraints due to the expressivity of the computer language. The main drawback of such methods is the lack of realism for generating motion, in particular for complex behaviors requiring the synchronization of multiple body parts.

2.2 Virtual Signers

Sensibly, signs differ from other communicative gestures, given the strict linguistic nature of their movements. They are indeed dependent on syntactic, semantic, and morphological constraints, as well as phonetic characteristics that encode the spatial features of signs conveyed by different channels (i.e., the gestures of the two arms and the two hands, facial expressions, and gaze direction). During the last decade, 3D virtual characters called virtual signers have been designed to provide increased accessibility for deaf people on a range of computing devices. Moreover, these avatars have given rise to different applications, including sign production, translation from text, and evaluation of sign synthesis that may support signed linguistics research.

Until now, signing avatars have been focused on generating sign utterances given a phonetic or phonological description of the sign sequence, using animation techniques as described above. With specific reference to French Sign Language, one of the first virtual signers was based on a description of the signing space, associated to a phonological description of hand-arm movements [Gibet et al. 2001]. Incorporating the HamNoSys [Prillwitz et al. 1989] sign language notation system as input, the ViSiCAST European project has designed an XML-based specification language called SigML [Kennaway 2003; Elliott et al. 2004], which can represent signing expressed in any signed language, and is interpreted into signed language gestures using a procedural animation technique. In the line of this project, the eSign project was undertaken as a response to the need for technologies that enable efficient production of sign language content over the Internet [Kennaway et al. 2007]. By using the SigML scripting notation and a client-side web browser plug-in to interpret this notation into motion data, a signing avatar can be incorporated in a variety of contexts. More recently, the Dicta-Sign project aims to develop the necessary technologies that make Web 2.0 interactions possible in different signed languages using high-level planners for signing avatars based on fine-grain geometrical descriptions [Delorme et al. 2009], or on knowledge-based descriptions [Fotinea et al. 2008].

Chiu et al. discuss a novel approach to translating from written Chinese to Taiwanese Sign Language, producing videos of signs using a bilingual corpus and sign data [Chiu et al. 2007]. In the ATLAS project, a virtual interpreter translates from Italian to Italian Sign Language (LIS). The system parametrizes pre-captured and hand-animated signs, to adapt them to the discourse context [Lombardo et al. 2010]. The user-based evaluation of American Sign Language generation has also been discussed in recent studies[Huenerfauth et al. 2007]. Important factors have been highlighted in this research, for example the influence of speed and pausing in animation of ASL [Huenerfauth 2009].

One of the main criticisms made by deaf signers regarding virtual agents is the lack of realism and expressiveness of avatars. Moreover, many avatar systems ignore the importance of facial expressions and gaze direction even though these components are crucial to comprehension in signed languages. Data-driven animation methods can be substituted for the above-discussed pure synthesis methods in order to improve the realism of produced gestures, making the avatar more expressive and human-like [Awad et al. 2009].

2.3 Data-Driven Methods for Embodied Communicative Agents

Data-driven methods constitute an appealing way to animate virtual avatars. Based on motion data, those methods also allow to modify the input data on purpose. Hence most of the previous work on data-driven animation methods present editing and composition techniques, with an emphasis on the re-use of motion chunks and the adaptation of captured motion for creating new motion sequences. The modification techniques involved are classical editing operations such as blending or concatenation as in Kovar et al. (2002), Liu and Popović (2002) or Mukai and Kuriyama (2005), but also spatial or temporal modifications as shown in Tak and Ko (2005) and Wang and Bodenheimer (2008). More recently, many data-driven approaches have also focused on building statistical models from motion data [Grochow et al. 2004; Chai and Hodgins 2007; Ikemoto et al. 2009]. Other relevant works include approaches that rely on qualitative annotations of motion clips [Arikan et al. 2003]. One can note that very few approaches deal with both motion-captured data and their implicit semantic content, and nearly nothing concerns communicative gestures. Stone et al. propose an approach for meaningfully synchronizing gesture and speech at common points of maximum emphasis [Stone et al. 2004]; in this kind of work, the entire body is controlled by motion capture. Another approach uses annotated videos of human behaviors to synchronize speech and gestures and a statistical model to extract specific gestural profiles: from textual input, a generation process then produces a gestural script which is interpreted by a motion simulation engine [Neff et al. 2008]. It should also be noted that motion capture data may be manipulated to enhance the expressivity of the gestures (e.g. [Wang et al. 2006]) or exagerate given traits.

As an added benefit, motion capture (mocap) data provides analytical material from which to extract specific features or parse generic features of signed languages, such as the dynamics of the movements or the spatial-temporal relationship between production channels, between kinematics and phonetics, etc. These invariants or user-dependent characteristics may be manually identified through an annotation process, or automatically computed through statistical or signal-processing methods, and re-incorporated into the data-driven animation techniques.

As a conclusion, the main challenge with respect to the state-of-the-art data-driven animation methods is *i)* to be able to synchronize and handle at the same time several modalities involved in communicative gestures, with different sources of data, to produce a continuous flow of animation *ii)* and to handle correctly, in terms of data structures, both the data and the associated semantic. Our work constitutes an original step in this direction. We know present the SignCom project, which motivated our animation system.

2.4 The *SignCom* Project

The *Signcom* project aims to improve the quality of real-time interaction between humans and virtual characters conversing with each other in French Sign Language. The results of this research will be valuable for the creation of intelligent and expressive interfaces for people who use signed languages. Three aspects of the interaction are studied in this project: the recognition of signs made by a user, the dialog which provides an adapted response, and the synthesis of an appropriate response by a virtual signer. The recognition system, in tandem with the dialogue generator, is able to progressively specify the entities of the discourse and their relations, with the possibility of confirming or canceling the transmitted messages or of raising ambiguities if necessary. Dialogue is processed in real time and in a restricted applied context with a limited vocabulary, allowing us to build new utterances from signs contained in the database.

This rest of the paper focuses on the synthesis part of the *Signcom* project, i.e., our fully data-driven virtual signer. We present a multichannel animation framework decomposed along different channels that represent information-conveying body parts: lower body, torso, arms, hands, head, face, and gaze.

The functional organization of *SignCom* is represented in Figure 1. The system is composed of two large building blocks: one, operating off-line, is a dually-indexed database containing both motion capture and semantic data, and the other, operating on-line, comprises the automatic recognition of signs, the animation of a virtual signer, and a go-between module that produces meaningful and appropriate dialog. The indexed database has been previously discussed in [Awad et al. 2009], in a manner similar to [Arikan et al. 2003]; the multimodal aspects of the animation have since been added.

3. CORPUS AND METHODOLOGY

3.1 Understanding Signed Language Motion

The notion of decomposing signs into various components is not new to the linguistic community. In 1960, William Stokoe debuted his system of *Tab* (location), *Dez* (handshape), and *Sig* (movement) specifiers that were to describe any sign [Stokoe 2005]. Since then, other linguists have expanded on Stokoe's decompositional system, introducing wrist orientation, syllabic patterning, etc. [Brentari 1999; Johnson and Liddell 2009].

However, signed languages are not restricted to conveying meaning via the configuration and motion of the hand; instead, they require the simultaneous use of both manual and non-manual components. The manual components of signed language include hand configuration, orientation, and placement or movement, expressed in the signing space (the physical three-dimensional space in which the signs are performed), while non-manual components consist of the posture of the upper torso, head orientation, facial expression, and gaze direction.

3.2 *SignCom* Corpus Development

With the above understanding of signed languages in mind, we and other members of the *SignCom* project specially constructed a corpus of signs to record for later use with the signing avatar. We detail here some of the challenges posed by certain

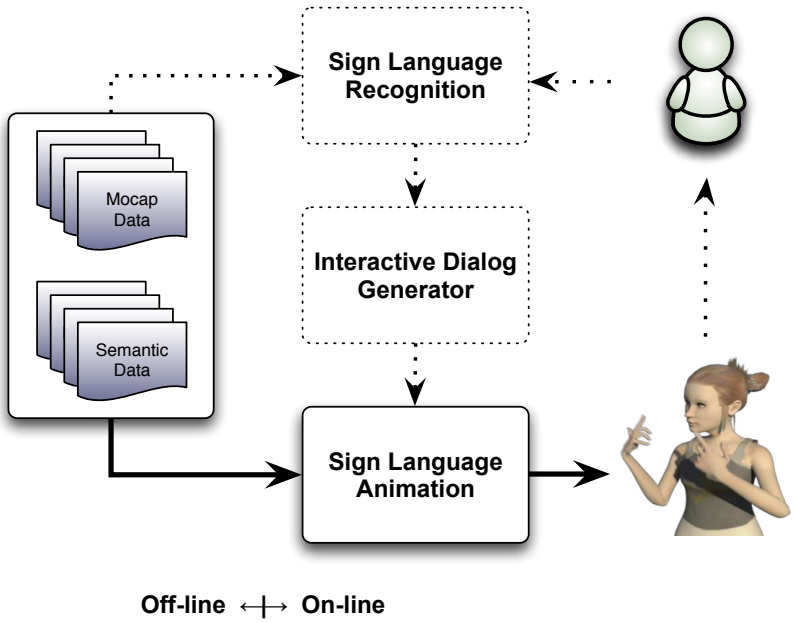

Fig. 1. An overview of the *SignCom* system. Only modules and data flows indicated in solid lines will be discussed in this paper.

aspects of LSF, and how we have chosen to incorporate such challenging data into our project experiments.

Spatial content. As signed languages are by nature spatial languages, forming sign strings requires a signer to understand a set of spatial-temporal grammatical rules and inflection processes. These processes have shaped the range of LSF signs recorded for the project.

The corpus was designed by a team of researchers that includes linguists, motion capture engineers, and computer scientists, among them both Hearing and Deaf [Duarte and Gibet 2010]. We chose a set of nouns, as well as depicting and indicating verbs which are modulated in the context of dialog situations. In one example, the sign INVITE can be modified grammatically to be understood as "I invite you", "you invite him", etc. Verbs that have typically been labelled *classifiers* and *size and shape specifiers* were also included.

Moreover, for the purposes of signed language synthesis, signing avatar corpora require many repetitions of the same sign in different contexts. This allows for the composition of new utterances with different components associated to different spatial locations. The repetition of signs also provides an important base for motion retrieval and re-use for animation purposes. With multiple phonological instances of the same sign recorded, a computer animator can choose a best-fit sign out of many, instead of forcing a single instance of the sign into a unique context.

Spatial-temporal aspects of hand movements. With signed languages being natural languages composed of spatial-temporal components, the question of the timing and dynamics of signs is critical. Specifically, the hand movements involved in a sign must be synchronized according to human motion principles and linguistic constraints in order for them to be believable.

The LSF sign DISAGREE is a compound sign formed from the signs THINK and DIFFERENT. In the sign, the *strong*[2] index finger moves from touching the center of forehead to being touched by the *weak* index finger in front of the forehead; the hands then move away from the center of the body while rotating outwards. Here, the motion of the weak hand is clearly synchronized to the strong hand, anticipating the arrival of the strong hand in front of the forehead. In other signs, we observe symmetrical or alternate roles of the two hands, with symmetry being usually defined about one of the three planes (sagittal, frontal and horizontal). In this case, both hand movements are synchronized.

As any component of a sign may modify the syntactic or semantic content, we should be able to acquire some knowledge of the temporal schemes characterizing the formation of signs along the different channels. For example, it is commonly understood that the handshape is attained before the beginning of the hand movement. Other relationships between left and right hand, between hand movement and facial expression, etc., should be identified and then utilized in a compositional animation system.

Synchronization is an important consideration for a signing avatar system, and ultimately would be a good candidate for automation. At this point, however, channel synchronization is performed by a member of the animation team.

Hand motion and handshape precision. Comprehension of signs requires accuracy in their formation. Particularly in fingerspelling, where each letter of an alphabet is named with a sign, the degree of openness of the fingers can be the sole differentiating factor between letters. Some handshapes differ only by the position of one finger or by whether or not it contacts another part of the hand. This calls for notable accuracy in the motion capture and data animation processes.

Non-manual components. While much of our description focuses on hand configuration and motion, important non-manual components are also taken into account, such as shoulder motions, head swinging, changes in gaze, or facial mimics. For example, eye gaze can be used to recall a particular object in the signing space; it can also be necessary to the comprehension of a sign, as in READ(v), where the eyes follow the motion of fingers as in reading. In the case of facial mimics, some facial expressions may serve as adjectives (i.e., inflated cheeks make an object large or cumbersome, while squinted eyes make it thin) or indicate whether the sentence is a question (raised eyebrows) or a command (frowning). It is therefore very important to preserve this information during facial animation.

[2]According to Johnson and Liddell, the strong hand is the hand used most actively during signers, which for most right-handed people is the right hand; the weak hand would thus be their left hand. This can be reversed for left-handed signers or when experienced signers sign two signs at the same time.

Fig. 2. Left, our native signer poses with motion capture sensors on her face and hands; right, our virtual signer in a different pose.

3.3 Data Conditioning and Annotation

The motion capture system used to capture our data employed Vicon MX infrared camera technology at frame rates of 100 Hz. The setup was as follows: 12 motion capture cameras, 43 facial markers, 43 body markers, and 12 hand markers. The photo at left of Figure 2 shows our signer in the motion capture session, and at right we show the resulting virtual signer.

In order to replay a complete animation and have motion capture data available for analysis, several post-processing operations are necessary. First, finger motion was reconstructed by inverse kinematics, since only the fingers' end positions were recorded. In order to animate the face, cross-mapping of facial motion capture data and blendshape parameters was performed [Deng et al. 2006]. This technique allows us to animate the face directly from the raw motion capture data once a mapping pattern has been learned. Finally, since no eye gazes were recorded during the informant's performance, an automatic eye gaze animation system was designed.

We also annotated the corpus, identifying each sign type found in the mocap data with a unique gloss so that each token of a single type can be easily compared. Other annotations follow a multi-tier template which includes a phonetic description of the signs [Johnson and Liddell 2009], and their grammatical class [Johnston 1998]. These phonetic and grammatical formalisms may be adapted to any sign language and therefore the multimodal animation system, which uses a scripting language based on such linguistics models, can be used for other sign language corpora and motion databases.

3.4 Multichannel Signed Language Data Composition

Our goal is to be able to produce new utterances from the corpus data by combining several channels, as depicted in Figure 3.

The sign composition follows the description of signs into manual and non-manual components (Section 3.1) along different channels phonetically annotated (Section 3.3). This composition process serves as inspiration for our animation system, though our goals do not require the phonetic specificity that linguists generally desire. Just the same, we must still encode how the multiple parts of the signer's

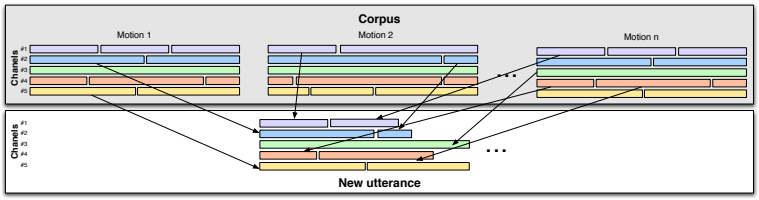

Fig. 3. Creating new utterances from the corpus. Annotation segments the data into multiple channels, and a new utterance is composed with several elements from the corpus.

body (*channels*) are articulated in parallel over time, and we must also specify motion along the channels that correspond to the linguistic categories of handshape, location, movement, orientation, facial expression, and gaze direction, as illustrated in Figure 4.

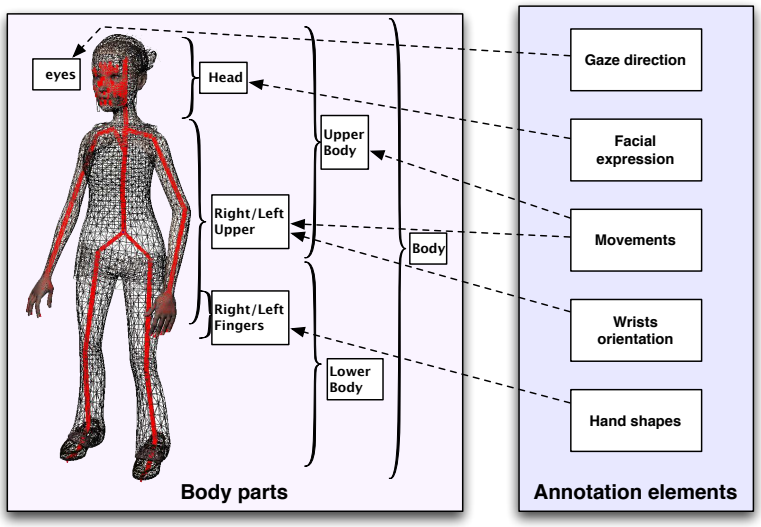

Fig. 4. The channels we manipulate in our animation system are inspired by the mechanics of signed languages, as proposed by signed language linguists. This figure shows the rough correlation between linguistic features and the channels of the *SignCom* animation system.

Despite there being some large technical differences between the semantic and phonological channels that we must consider for this work, their general correspondence aids in making this new work more manageable. This approach thus follows Vogler and Metaxas's modeling of the simultaneous features of ASL into independent channels for recognition purposes [Vogler and Metaxas 2004].

Before proceeding, we note that the choice of motion elements to be combined, being a linguistic issue, is governed in this paper by simple semantic and grammatical rules, involving motion segments on independent channels that have specific meanings (hand movements or handshapes). In our experiments, our different composition scenarios were carefully designed by a signed language linguist, and

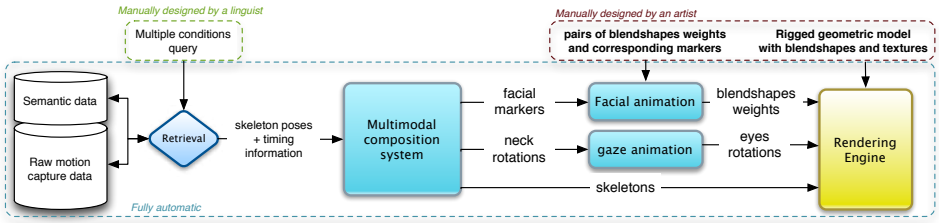

Fig. 5. Overview of the animation system.

expressed as a simple script used by the animation engine. Our more pressing aim is to show the feasibility, from an animation point of view, of a multichannel compositional system that integrates the inherent constraints of signed language utterances.

4. OUR DATA-DRIVEN ANIMATION SYSTEM

We will now describe our animation system and how we arrive at comprehensible signed language sequences such as those used in the evaluation we discuss in Section 5. An overview of the system is given in Figure 5. The process begins with a list of motion elements paired with timing information, retrieved from two different databases that contain semantic (annotation) and raw (motion capture) data (Section 4.2). Then our multichannel composition system builds a new motion expressed as a sequence of skeletal postures (Section 4.3). These postures contain information that encodes body and hand configurations as well as facial markers. Next, the facial markers are turned into a new geometric facial configuration by means of blendshapes and a learning method (Section 4.4); eye animation is also inferred from this skeletal posture (Section 4.5). Finally, the rendering engine computes the final avatar image.

4.1 Body Part Nomenclature

As discussed in the section on annotation above, we divide the skeleton into several sub-articulated chains. Throughout the rest of this paper, we will refer to these chains as *body parts*, and we associate each body part to a channel. The animation system functions off of the body parts shown in Figure 4, labeled *UpperBody*, *Spine*, *Head*, *RightUpper*, *RightFingers*, *LeftUpper*, *LeftFingers*, and *LowerBody*.

4.2 Data Coding and Retrieval

As shown at the beginning of this paper in Figure 1, the *SignCom* interaction system is divided into two parts: an off-line process of data storage and on-line data retrieval for real-time interaction. The originality of the work presented here originates in the methodology used for data storage and in the streaming method used to retrieve motion data. Our system provides fast and efficient motion retrieval during the animation process, taking into consideration the spatial and temporal aspects of signed language motion described above. The nature of the different types of information encoded in and by signs makes it necessary to store data in two different structures, namely a semantic database for textual annotations, and

a raw database for motion capture data.

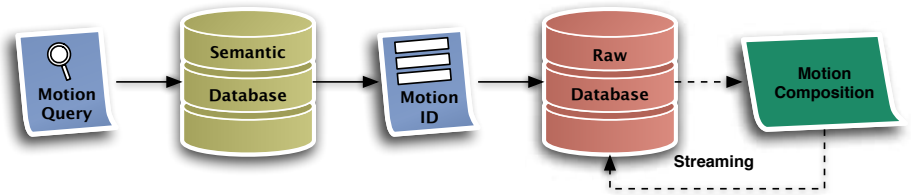

Fig. 6. Data retrieval and stream loading system. The semantic database, containing textual information from the annotation process, is queried first. The motion data corresponding to the obtained results are then streamed to the motion composition process.

As depicted in Figure 6, retrieving data from the databases is divided into two parts. The first part of the process consists of querying the semantic database, allowing us to extract data corresponding to a list of MotionIDs. Briefly, these MotionIDs represent the canonical index data structure of a motion element. Each contains the name of the sequence in which the chunk occurs, time stamps relative to the beginning of this sequence (noted as Frame In and Frame Out), and the involved body parts. This mapping between the annotation and motion data constitutes the semantic database (Figure 7), which is automatically constructed from an XML hierarchical description language provided by the annotation tool (ELAN in our case). We emphasize here the one-to-many nature of this mapping, where any one gloss from the textual annotation can be associated with several different instances of the same gesture. As one example, the gloss COCKTAIL in Figure 7 corresponds to two MotionIDs, 1 and 4.

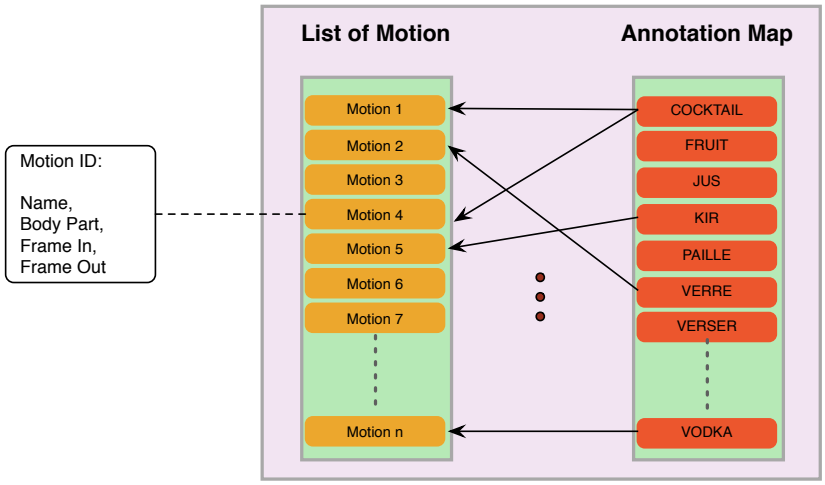

Fig. 7. The semantic database is a one-to-many mapping between annotated glosses and MotionIDs, which are canonical index data structures of motion elements.

In our application, retrieving data from the semantic database is achieved by specifying multiple-condition queries, the conditions of which can be keywords and/or body parts, and which return one or several MotionIDs. Secondarily, the query results are interpreted so that each MotionID leads to accessing the raw database and rendering the corresponding motion frames.

Raw motion database. In our system, the internal representation of a motion contains an association of the hierarchical structure (commonly called a bindpose), and a list of relative transformations for each joint. The transformation for the root joint contains joint position and rotation (expressed in quaternions), while the transformation for the rest of the joints contains only a rotation. The time needed to read a motion file into this internal representation depends naturally on the complexity of the parser and the amount of geometrical computations, and is usually far from being negligible, preventing dynamic loads in our interactive application. Motion files are thus loaded and interpreted one time, and stored as a sequence of bits in our database, having written our own serialization process for this purpose.

Traditional databases function with a set of pair-valued data: one key (preferably unique) is associated to the useful data (in our case the motion). The simplest way to proceed is to associate for instance the whole motion file with a unique key, which might be defined as the name of the original data file. The whole sequence is then handled by the database manager, and stored on the hard drive. This approach assumes that when retrieving the motion, all the data will be reconstructed in the CPU memory. In the context of a real-time animation controller, where small pieces of the motion are dynamically combined to achieve a desired goal, this approach is no longer efficient. We have thus designed our database to handle a different

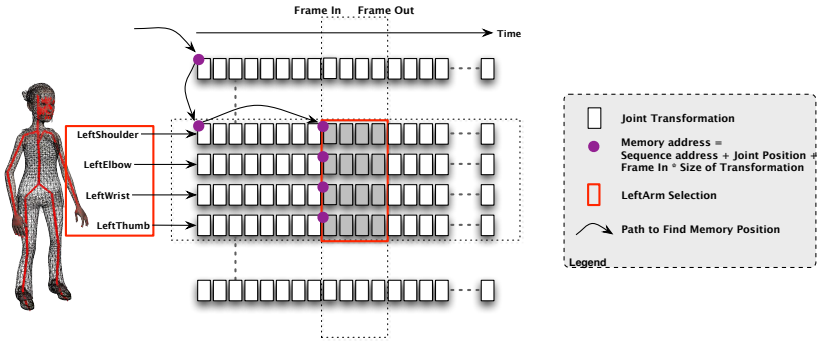

Fig. 8. Storage and data access in the raw database

Table I. Querying the Databases

	Semantic Query(ms)	Number of Frames	Size of Motion motion (ko)	Complete loading(ms)	Streaming access by 256ko fragment (ms)
gloss SALADE	0.0652	64	426.5	9,79	3.533
A scenario	//	5739	38 244.7	3 785	3.533

data representation, allowing us to retrieve any part of a motion corresponding to a given annotation element (Figure 8).

Decomposing motions in the database is innovative because only a small portion of the motion (associated to a query result) is reconstructed in the memory. However, in traditional databases, data decomposition generally yields an increasing number of entries, which usually increases the search time and the index size. Yet in our case we consider each motion to be a list of transformations with given sizes; therefore it is easy to find the memory address of a list of transformations as a linear combination of the sequence address, joint offset, and time stamps, as illustrated as a path in Figure 8.

To complete the access to the raw motion capture data, we have developed a streaming system which loads the motion to animate in a fragment-by-fragment manner during the animation process (a fragment being a small set of transformations), and with regards to the need of the motion composition system. This avoids costly access to large elements which could result in a drop in frame rate during the execution of the application, and gives the process a small memory footprint. Computationally, this allows the interactive nature of this animation system to move forward, since database search and data load time become negligible during animation. As examples, results for different queries are shown in Table I.

4.3 Motion Composition

From our corpus of mocap data, our animation system computes a skeleton using a pre-defined morphology of joints and bindposes, which can be represented hierarchically as a tree of joints or articulations. Within the skeleton, we have identified sub-skeletons composed of potentially non-exclusive subsets of joints, including the upper body, lower body, arms, hands, head, etc. A controller associated to each

sub-skeleton can set the system in motion using different techniques, i.e., motion playback, keyframe interpolation, inverse kinematics, etc.

The motion composition process can be divided into spatial and temporal composition processes. The spatial composition process uses motions computed for each controller's sub-skeleton, combining them in a priority scheme that depends on the desired animation; generally, the smaller sub-skeletons have a higher priority level, as shown in Figure 9. Temporal composition occurs for the set of controllers attached to the skeletal elements. Each controller has its own timing interval and a playback style (e.g., play once, repeat, reverse, etc.), and the blender process is responsible for blending the motions.

Figure 9 is a graphical representation of how we organize blenders and controllers during composition. Algorithm 1 shows how the blender controllers blend sub-skeletons both temporally and spatially. Finally, we have developed a simple script language in order to easily specify different animation scenarios, containing controller and blender information associated with time stamps.

The controllers applied on sub-skeletons (body, arms, hands, torso, etc.) are traditional controllers that are not described in this paper. More specific controllers developed for facial and eye animation are described in the rest of the section. Both of them use motion captured data as input and produce information useful for the animation engine.

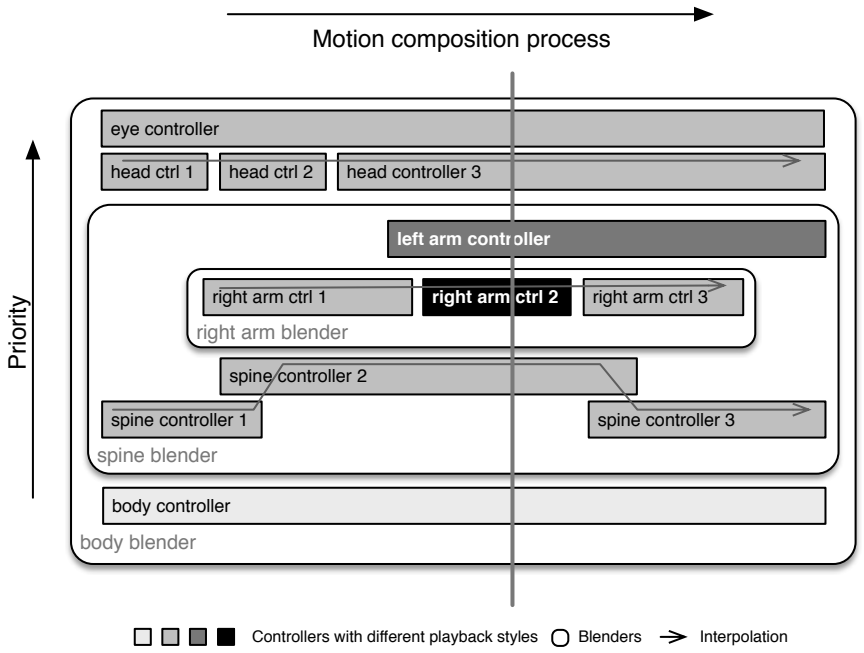

Fig. 9. Blenders are arranged hierarchically in the system and contain a series of controllers to animated different sections of the body. Skeletons are computed according to priority of controllers and, over time, the engine produces a stream of fluid motion.

Algorithm 1 Blending Algorithm

```
for all joints do
  set joint_weights to 1 and joint_transformations to Identity
end for
while childController && a joint weight > 0 do
  if time ∈ [startTime - fadeIn, endTime + fadeOut] then
    find weight w wrt. time and compute controller skeleton
    for j = 0 to skeleton.jointCount do
      k = joint index in joint_transformations
      if joint_weights[k] > 0 then
        if joint_weights[k]! = 1 then
          joint_transformations[k] = interpolate wrt. w between
          joint_transformations[k] and skeleton.joint_transformations[k]
        else
          joint_transformations[k] = skeleton.joint_transformations[k]
        end if
        joint_weights[k] = max(0, joint_weights[k] - w)
      end if
    end for
  end if
end while
```

4.4 Facial Animation

Facial animation by blendshapes is a popular technique in the animation community, and we have chosen likewise. Following this method, the animation system blends several key facial configurations, manually designed by an animator, to produce appropriate facial animations. In order to choose the blending weights at each moment, the system uses the facial mocap data contained in the currently-processed skeleton, as described below.

Cross-mapping of facial mocap data and blendshape parameters. The process of cross-mapping mocap data and blendshapes parameters can be problematic for the animation process: it is often challenging to quantify the relation between facial mocap data and the animation parameters of a blendshape. Traditional approaches to solving this problem identify pairs of mocap data and blendshape parameters that are carefully selected and designed by the animator [Deng et al. 2006]. These pairs are then used in a learning process that determines the selection of corresponding blendshape parameters from new mocap data input values. Other current methods largely rely on radial basis functions and kernel regression to achieve these steps [Cao et al. 2005; Deng et al. 2006; Deng et al. 2006; Liu et al. 2008].

However, such methods have several drawbacks: a number of localized basis functions have to be chosen prior to the learning process, and the result is conditioned by the quality and density of input data. Thus, noisy input often yield bad estimates, this being known as the classical over-fitting problem.

In our work, both the body and facial data were recorded at the same time, and

the positions of the facial markers in particular were observed to be quite noisy, resulting in marker inversions. For these reasons, we consider the problem as a probabilistic (Bayesian) inference problem and use a separate learning technique based on Gaussian Process Regression, which is well-known in the machine learning and vision communities [Rasmussen and Williams 2005].

Gaussian Process Regression (GPR). The GPR approach aims to solve the following prediction problem: given p observations $\mathbf{X} = (X(l_1), \ldots, X(l_p))^T$ localized at the l_i sites, one looks at the estimation of $X(l_k)$ at a given unobserved localization l_k. This problem is solved by assuming that the underlying generative process is Gaussian, and by building the conditional distribution $p(X(l_k)|\mathbf{X})$ which is itself Gaussian.

In our approach, unknown sites correspond to new facial marker configurations (as produced by the previously-described composition process), and the corresponding estimated value is a vector of blendshape weights. Since the dimensions of the learning data are rather large (123 for marker data and 50 for the total amount of blendshapes in the geometric model we used), we rely on an online approximation method of the distribution that allows for a sparse representation of the posterior distribution [Csató and Opper 2002]. As a preprocess, facial data is expressed in a common frame that varies minimally with respect to face deformations. The upper-nose point works well as a fixed point relative to which the positions of the other markers can be expressed. Secondly, both facial mocap data and blendshape parameters were reduced and centered before the learning process.

Figure 10 shows an illustration of the resulting blended faces along with the different markers used for capture.

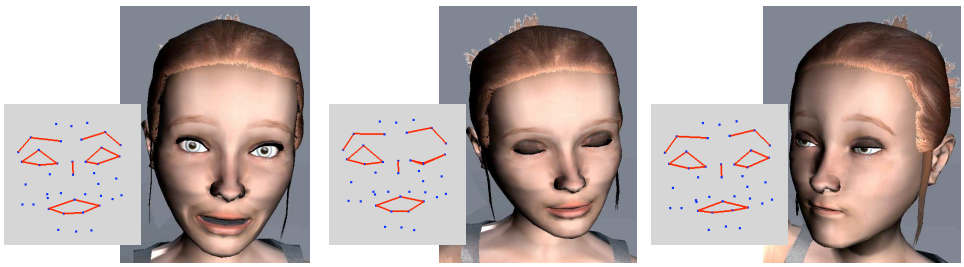

Fig. 10. Results of the facial animation system. Some examples of faces are shown, along with the corresponding markers position projected in 2D space.

4.5 Eye Animation

Our capture protocol was not able to capture the eye movements of the signer, even though it is well-known that the gaze is an important factor of non-verbal communication and is of assumed importance to signed languages. Recent approaches to model this problem rely on statistical models that try to capture the gaze-head coupling [Lee et al. 2002; Ma and Deng 2009]. However, those methods only work for a limited range of situations and are not adapted to our production pipeline. Other approaches, like the one of Gu and Badler [Gu and Badler 2006], provide a

computational model to predict visual attention. Our method follows the same line as we use a heuristic synthesis model that takes the neck's motion as produced by the composition process as input and generates eye gazes accordingly. First, from the angular velocities of the neck, visual targets are inferred by selecting times when the velocity passes below a given threshold for a given time period. Gazes are then generated according to those targets such that eye motions anticipate neck motion by a few milliseconds [Warabi 1977]. This anticipatory mechanism provides a baseline for eye motions, to which glances towards the interlocutor (camera) are added whenever the neck remains stable for a given period of time. This ad-hoc model thus integrates both physiological aspects (modeling of the vestibulo-ocular reflex) and communication elements (glances) by the signer. Figure 11 shows two examples of eye gazes generated by our approach. However, this simple computational model fails to reproduce some functional aspects of the gaze in signed languages, such as referencing elements in the signing space. As suggested in the following evaluation, this factor was not critical with regards to the overall comprehension and believability of our avatar, but can be an area of enhancement in the next version of our model.

Fig. 11. The two types of glances produced by our system (a) direct look to the interlocutor (b) anticipation of the neck rotation

5. EVALUATION OF THE *SIGNCOM* DATA-DRIVEN SIGNING AVATAR

Evaluation of virtual signers [Huenerfauth et al. 2007] has been performed in the USA, using a native ASL signer whose movements were recorded by motion capture techniques (using datagloves and a mocap suit), but with only hand and body movements being recorded. To our knowledge, using simultaneous motion capture for hand and body movements, facial expression, and gaze direction has not yet been performed, therefore we have devised an evaluation of our system which is divided in two parts: first, we test the importance of facial expressions and gaze movements when generating LSF sentences, and second, we evaluate the animation system's ability to produce a realistic and comprehensible signing avatar.

We test these abilities by showing LSF users videos of the avatar in action, showing pairs of videos synthesized from motion capture data (with channel/without channel), and comparing native LSF movements (from motion capture data) to synthesized ones (from reconstructed sequences).

5.1 Survey Dissemination and Respondents

Evaluation was carried out via a web-based survey, sent to members of the French Deaf community who were requested to disseminate the link. In all, 38 people completed the survey to its end, and we consider 25 respondents who self-reported to know French Sign Language (LSF) at a *bon* (good, N=8), *très bon* (very good, N=6), or *signeur natif/expert* (native signer/expert, N=11) level. Respondents

were mostly from the Paris area (N=16), and were most often female (N=18). We received responses from a range of ages (19-56), and had a mix of hearing (N=8) and deaf (N=17) respondents.

5.2 Facial Expression Tests

In the French signed language linguistics community, there is a relatively heavy emphasis placed on the role of facial expressions in the use of LSF, and rightly so since facial expressions encode both grammatical (adjectival) and prosodic/pragmatic details. As a result of this emphasis, Deaf French avatar users demand appropriate expressivity in the avatar's face.

To support the aims and outcomes of *SignCom*, we carried out three facial expression tests in order to better grasp respondents' comprehension of signing avatar facial expressions in general, and to see if our method for animating the face using blendshapes was sufficient.

Three videos were prepared with three different sign sequences, and each with a different method for animating facial expressions:

A manually-synthesized facial animations were created, using facial blendshapes determined by an animator; this sequence's motion capture file was replayed along with the body motion, and used as a baseline LSF native animation sequence

B the data-driven synthesized facial animation method described in Section 4.4 uses the facial mocap data and blendshape parameters learned from sequence A to animate this sequence

C the face was left unanimated during the sign sequence

Survey takers were shown three pages, each containing a pair of videos; the same set of questions on each page asked respondents which video they preferred and why.

When comparing the manual and data-driven animated facial expressions in sequences A and B, respondents generally had no preference, choosing instead to comment on the quality of the avatar's signing or other topics. As a group, the respondents tended slightly toward sequence B, the sequence with the data-driven-synthesized facial expressions. On a scale of -2 – 2 with -2 being a strong preference for sequence B and 2 being a strong preference for sequence A, the responses averaged $\bar{x} = -.44, \sigma = 1.26$.

Indeed, the qualitative responses during this pairing made it clear that respondents didn't notice much of a difference between the two facial animations. Instead, participants took the opportunity to describe other suggestions they had to improve the animation system (detailed later in this section). We consider this finding to validate the use of blendshapes for the facial animation process for signing avatars, even when these blendshapes are taken from a single motion capture sequence and applied to others.

In comparing videos A and B with video C, respondents trended towards the videos with facial expressions, though not as strongly as we would have predicted. For example, when -2 meant a strong preference for video C and 2 meant a strong preference for video A respondents rated the pairing $\bar{x} = .52, \sigma = 1.48$, and when -2 meant a strong preference for video C and 2 meant a strong preference for video

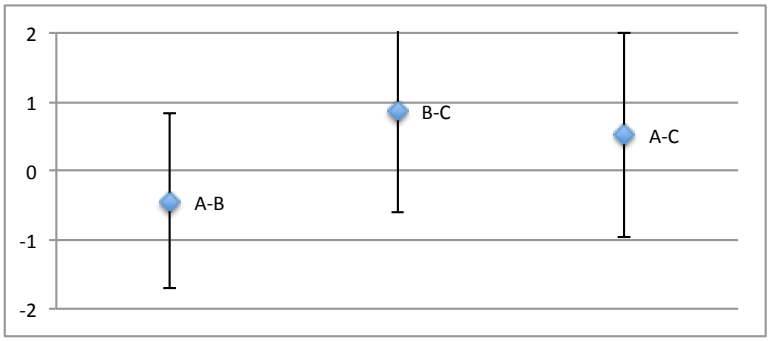

Fig. 12. Ratings of each of the three facial expression tests. We suspected that users would find an order of naturalness that would follow A > B > C, however this trend was not always followed.

B respondents rated the pairing $\bar{x} = .88, \sigma = 1.48$. Results for all three tests are summarized graphically in Figure 12.

During these two tasks, we received a number of comments preferring the facial expressions in videos A and B over those in video C, in contrast to the lack of comments we received in the first task suggesting that both videos from the first task were accessible to the respondents. This again supports our theory that using facial blendshapes to animate the face not only saves on computation cost, but also provides a convincing and preferable facial animation for users of signing avatars.

5.3 Gaze Direction Test

Referring again to the French signed language linguistics community, some have argued that signers use their eye gaze for prosodic or even grammatical functions during the discourse. Thus we animated two almost identical videos, the difference between them being the animation of the eyes: in one video the signer's eyes moved as described in Section 4.5, and in the other video the signer's eyes were fixed.

The respondents generally had no preference for either video; 20 of them (80%) rated the pair a 0 on a -2 – 2 scale. Qualitative responses verified that respondents didn't notice the real difference between the two videos, with some believing that we had altered the coloring on the avatar and others using the comment box to suggest other non-eye-related improvements. This suggests that eye animation is rather unimportant in the overall task of understanding a signing avatar, simply because the eyes were too small in the video to notice a difference, or because deficits in other channels supersede the importance of eye animation.

5.4 Motion Capture Playback and Constructed Sequence Tests

The first video shown to respondents was a control video of a simple replay of a motion capture sequence. In the sequence, the avatar explains to the audience that she has recently held a cocktail party for her friends, and describes the preparations she made for them to come over.

We asked respondents to rate their comprehension of the signs used, their comprehension of the entire story, and the realism of the avatar using a 1 – 5 Likert scale, with 5 being very realistic or very comprehensible. On average, respondents

rated the motion capture sequence as 3.12 for realism, 3.48 for sign comprehension, and 3.88 for story comprehension, as shown in Table II. While these numbers were lower than expected, they still remain above the median threshold of 3.

The second and third tests in this section showed two constructed dialogues and asked respondents to rate them for the same three factors as in the control playback sequence (realism, sign comprehension, and story comprehension). This allowed us to test our system's ability to produce convincing linguistic utterances by combining motion segments across different channels.

The two constructed dialogues we used contain a large number of tokens related to the cocktail party scenario we recorded while building our corpus. With a large variety and frequency of cocktail-related lexemes in our corpus, we are able to produce a number of novel utterances around the same subject. One of these constructed sequences is transcribed below, and diagramed to show motion retrieval and combination in Fig 13.

> I asked my friend, "what do you want?"
> (S)he said, "Well, I don't like orange juice. What would you suggest?"
> "I'd suggest a Cuba Libre," I responded.

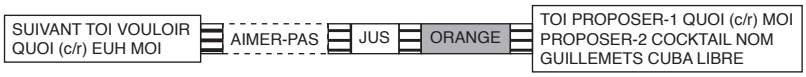

Fig. 13. Signs can be rearranged to create novel phrases. Here, signs are retrieved from two different recording takes (white and gray backgrounds) and linked with transitions created by the animation engine (striped background). The sign AIMER (*like*) is reversed to create AIMER-PAS (*don't like*). The signs shown here represent the manual animation of the avatar during a single sequence; other tracks are animated simultaneously to move the body and head in ways meaningful to the discourse.

Respondents rated stories 1 and 2 similarly to the control playback sequence. This suggests that although those surveyed have hesitations about signing avatars,

Fig. 14. Animation strip of the first scenario.

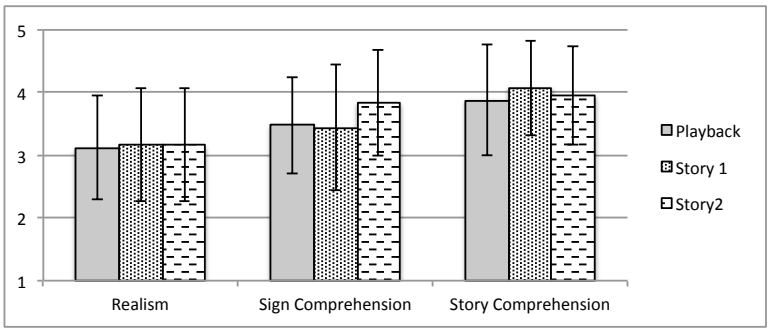

Fig. 15. Ratings of each of the three sequences grouped by question. There is obvious visual similarity to the responses, which has been confirmed with t-test P values.

they don't find our concatenated sequences any less real or understandable than simple playback sequences. Responses are quantified in Table II, and grouped by question in Figure 15.

Table II. Ratings given by respondents for a simple playback sequence and two stories created with our animation engine. The response scale ranged from 1 as not at all realistic/comprehensible to 5 as very realistic/comprehensible. P values compare the story sequences to the control playback sequence.

	Playback ($\bar{x}$, σ)	Story 1 ($\bar{x}$, σ; P)	Story 2 ($\bar{x}$, σ; P)
Realism	3.12, .83	3.16, .90; .8709	3.16, .90; .8709
Sign Comprehension	3.48, .77	3.44, 1.00; .8748	3.84, .85; .1231
Story Comprehension	3.88, .88	4.08, .76; .3940	3.96, .79; .7367

Given the need for the human signer to wear distracting markers during recording sessions we were unable to evaluate concatenated sequences against videos of the human signer performing the same sequences. It is conceivable that different instantiations of the same phrase from different recording techniques (video, mocap playback, mocap concatenation) be evaluated against each other, but this is work for future studies.

5.5 Qualitative Results

As mentioned previously, respondents were asked to give written impressions about the avatar throughout the evaluation. These provide us with valuable feedback on issues arising in signing avatar use that we hadn't previously thought to focus on.

A sizable amount of respondents (N=12) state that sign comprehension was compromised by poor hand configurations on the part of the avatar. We suspect that our kinematic model for the hand was oversimplified: between the motion capture data and the specificities of the geometric model, finger contact does not often occur when expected. Clearly, this model has to be reconsidered in future versions of our virtual signer.

Regarding facial animation, some respondents found that our avatar lacked expressivity, which could be caused by two factors. Firstly, producing expressive

facial animations is a notoriously difficult problem, and the data-driven blendshape method is directly tied to the quality of the input data. Given that simultaneously capturing all the channels of a signer's performance is difficult, the resulting data is sometimes noisy or incomplete. Secondly, the artistic choice made by the artist while producing the blendshapes could influence facial expression realism and could be the subject of further evaluation.

Other comments (N=4) noted that the position of the virtual interlocutor (virtual camera position) must be precisely controlled to ensure that participants feel that the avatar is speaking directly to them, as opposed to above, below, or around them. Finally, some respondents (N=3) noted that the avatar was too skinny or had the impression of having had a facelift; such artistic choices should be rectified in future implementations so as not to detract from the interaction.

6. DISCUSSION AND CONCLUSIONS

Our system is able to work with a multichannel representation of signed language, and produce real-time data-driven animations that include body, hand, face, and eye motion. Importantly, it is able to do so with realism and comprehensibility - similar to that of replayed motion capture sequences.

We have detailed a preliminary evaluation of our techniques for generating utterances in LSF. Several results have been highlighted: first of all, our experiments have confirmed the importance of including facial expressions in signed language animations; furthermore, our facial animation method exhibits strong qualities, such that there was no significant difference between our method and manually-synthesized animations (i.e., those performed by an animator). This result reinforces the idea of using such a data-driven model for synthesizing any facial expression from motion capture data. Surprisingly, we also concluded that in the experiments conducted with our signing avatar, gaze direction had no particular significance on the sign stream. Finally, our motion composition process allows us to form novel utterances for which the results are promising, as respondents to our survey were unable to dissociate the synthetic motion from the playback movements.

Although we are convinced that the framework reported in this paper is one of the most advanced attempts to produce utterances by a virtual signer that are believable and better accepted by the deaf people, several unresolved problems must still be addressed in our data-driven techniques.

Signs are in effect gestures that require extreme precision and rapidity in the acquisition process; as a result, imperfect sign formation or improper sign synchronization can alter the semantic content of an utterance. Therefore, an expressive data-based animation system should handle all the spatial inflections and timing variations which are due to coarticulation effects. This is feasible if the basic motion chunks used to reconstruct the signing sequence are whole signs or glosses. It is clear however that extracting and using the separate phonetic components which should be adapted and brought together to form complete signs remains difficult.

The appearance of the virtual signer may also have an impact on the believability of the avatar. While realistic avatar could fail in the well known problem of the

"uncanny valley", other representations of the avatar (with a more abstract shape or a cartoonish representation) could also be tested.

In the near future, we hope to make significant improvements in our animation system:

i by introducing new controllers that are able to handle different constraints (e.g., spatial coherency, collision detection, or dynamics of movements),

ii by developing other ways of combining partial human motions that account for the coordination schemes between various channels, and

iii by developing new evaluation methodologies that more thoroughly analyze the degree of comprehension of the signs (i.e., through more detailed questionnaires), and the degree of expressivity of our virtual signer.

ACKNOWLEDGMENTS

This work is part of the *SignCom* project, supported as an Audiovisual and Multimedia project by the French National Research Agency (ANR).

REFERENCES

Arikan, O., Forsyth, D. A., and O'Brien, J. F. 2003. Motion synthesis from annotations. *ACM Transactions on Graphics 22,* 3 (July), 402–08.

Awad, C., Courty, N., Duarte, K., Le Naour, T., and Gibet, S. 2009. A combined semantic and motion capture database for real-time sign language synthesis. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*. Lecture Notes in Artificial Intelligence, vol. 5773. Springer-Verlag, Berlin, Heidelberg, 432–38.

Brentari, D. 1999. *A Prosodic Model of Sign Language Phonology*. MIT Press, Cambridge, MA.

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. 2005. Expressive speech-driven facial animation. *ACM Transactions on Graphics 24,* 4 (October), 1283–302.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. F. 2000. *Embodied Conversational Agents*. The MIT Press.

Chai, J. and Hodgins, J. 2007. Constraint-based motion optimization using a statistical dynamic model. *ACM Transactions on Graphics 26,* 3 (July), 686–696.

Chiu, Y., Wu, C., Su, H., and Cheng, C. 2007. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Trans. on Pattern Analysis and Machine Intelligence 29,* 1 (jan), 28–39.

Csató, L. and Opper, M. 2002. Sparse on-line gaussian processes. *Neural Computation 14,* 3 (March), 641–68.

Delorme, M., Filhol, M., and Braffort, A. 2009. An architecture for sign language synthesis. In *Proc. of Gesture Workshop 2009*. LNCS. Bielefeld, Germany.

Deng, Z., Chiang, P.-Y., Fox, P., and Newmann, U. 2006. Animating blendshape faces by cross-mapping motion capture data. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*. Redwood City, California, 43–48.

Deng, Z., Newmann, U., Lewis, J. P., Kim, T.-Y., Bulut, M., and Narayanan, S. 2006. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Transactions on Visualization and Computer Graphics 12,* 6 (November), 1523–34.

Duarte, K. and Gibet, S. 2010. Heterogeneous data sources for signed language analysis and synthesis: The signcom project. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (19-21). European Language Resources Association (ELRA), Valletta, Malta.

Elliott, R., Glauert, J., Jennings, V., and Kennaway, J. 2004. An overview of the sigml notation and sigml signing software system. In *Workshop on the Representation and Processing of Signed Languages, 4th Int'l Conf. on Language Resources and Evaluation*.

Fotinea, S., Efthimiou, E., Caridakis, G., and Karpouzis, K. 2008. A knowledge-based sign synthesis architecture. *Univ. Access Inf. Soc. 6*, 405–418.

Gibet, S., Lebourque, T., and Marteau, P. 2001. High level specification and animation of communicative gestures. *Journal of Visual Languages and Computing 12*, 657–687.

Grochow, K., Martin, S., Hertzmann, A., and Popovic, Z. 2004. Style-based inverse kinematics. *ACM Transactions on Graphics 23,* 3 (Aug.), 522–531.

Gu, E. and Badler, N. 2006. Visual attention and eye gaze during multiparty conversations with distractions. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*. Lecture Notes in Computer Science, vol. 4133. Springer-Verlag, Berlin, Heidelberg, 193–204.

Hartmann, B., Mancini, M., and Pelachaud2, C. 2006. Implementing expressive gesture synthesis for embodied conversational agents. *Lecture Notes in Computer Science : Gesture in Human-Computer Interaction and Simulation 3881/2006*, 188–199.

Huenerfauth, M. 2009. A linguistically motivated model for speed and pausing in animations of american sign language. *ACM Trans. Access. Comput. 2*, 9:1–9:31.

Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J. 2007. Evaluating american sign language generation through the participation of native asl signers. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. Assets '07. ACM, New York, NY, USA, 211–218.

Huenerfauth, M., Zhou, L., Gu, E., and Allbeck, J. 2007. Design and evaluation of an american sign language generator. In *45th Annual Meeting of the Association for Computational Linguistics. Workshop on Embodied Language Processing*. Prague, Czech Republic.

Héloir, A. and Kipp, M. 2010. Real-time animation of interactive agents: Specification and realization. *Applied Artificial Intelligence 24,* 6, 375–391.

Ikemoto, L., Arikan, O., and Forsyth, D. 2009. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics 28,* 1, 1–12.

Johnson, R. E. and Liddell, S. K. 2009. *Sign Language Phonetics: Architecture and Description*. Forthcoming.

Johnston, T. 1998. The lexical database of AUSLAN (Australian Sign Language). In *Proceedings of the First Intersign Workshop: Lexical Databases*. Hamburg.

Kendon, A. 1993. *Tools, Language and Cognition*. Cambridge University Press, Chapter Human gesture, 43–62.

Kennaway, J. R. 2003. Experience with, and requirements for, a gesture description language for synthetic animation. In *Proc. of Gesture Workshop 2003*. LNCS. Genova, Italy.

Kennaway, J. R., Glauert, J. R. W., and Zwitserlood, I. 2007. Providing signed content on the internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact. 14,* 3, 15.

Kipp, M., Neff, M., Kipp, K. H., and Albrecht4, I. 2007. Toward natural gesture synthesis: Evaluating gesture units in a data-driven approach. In *Intelligent Virtual Agents (IVA'07)*. 15–28.

Kita, S., van Gijn, I., and van der Hulst, H. 1997. Movement phase in signs and co-speech gestures, and their transcriptions by human coders. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*. Lecture Notes in Computer Science, vol. 1371. Springer-Verlag, London, 23–35.

Kopp, S., Krenn, B., Marsella, S., Marshall, N., Pelachad, C., Pirker, H., Thórisson, K., and Vilhjlmsson, H. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Proc. of Intelligent Virtual Agents, eds. J. G. Carbonell and J. Siekmann*. Marina Del Rey, USA, 205–217.

Kopp, S. and Wachsmuth, I. 2004. Synthesizing multimodal utterances for conversational agents. *Journal Computer Animation and Virtual Worlds 15(1)*, 39–52.

Kovar, L., Gleicher, M., and Pighin, F. 2002. Emotion from motion. In *Proc. of Int. Conf. on Computer Graphics and Interactive Techniques*. San Antonio, Texas, USA, 473–482.

Kranstedt, A., Kopp, S., and Wachsmuth, I. 2002. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents - let's specify and evaluate them*. Bologna, Italy.

Lee, S. P., Badler, J. B., and Badler, N. I. 2002. Eyes alive. *ACM Transactions on Graphics 21,* 3, 637–644.

Liu, C. K. and Popović, Z. 2002. Synthesis of complex dynamic character motion from simple animations. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. San Antonio, Texas, USA, 408–16.

Liu, X., Mao, T., Xia, S., Yu, Y., and Wang, Z. 2008. Facial animation by optimized blendshapes from motion capture data. *Computer Animation and Virtual Worlds 19,* 3-4 (September), 235–45.

Lombardo, V., Nunnari, F., and Damiano, R. 2010. A virtual interpreter for the italian sign language. In *IVA*. 201–207.

Ma, X. and Deng, Z. 2009. Natural eye motion synthesis by modeling gaze-head coupling. In *2009 IEEE Virtual Reality Conference*. Lafayette, Louisiana, USA, 143–50.

McNeill, D. 1992. *Hand and Mind - What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, IL.

Mukai, T. and Kuriyama, S. 2005. Geostatistical motion interpolation. *ACM Transactions on Graphics 24,* 3, 1062–1070.

Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics 27,* 1 (March), 233–51.

Noot, H. and Ruttkay, Z. 2005. Variations in gesturing and speech by gestyle. *Int. J. Hum.-Comput. Stud. 62,* 2, 211–229.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. 1989. *Hamburg Notation System for Sign Languages - An Introductory Guide*. University of Hamburg Press.

Rasmussen, C. E. and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning*. The MIT Press.

Stokoe, W. C. 2005. Sign language structure: an outline of the communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education 10,* 1, 3–37. Originally published as Studies in Linguistics, Occasional Papers 8 (1960), by the Department of Anthropology and Linguistics, University of Buffalo, Buffalo, NY.

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., and Bregler, C. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of ACM SIGGRAPH 2004*. ACM Transactions on Graphics, vol. 23. ACM, New York, NY, USA, 506–13.

Tak, S. and Ko, H.-S. 2005. A physically-based motion retargeting filter. *ACM Transactions on Graphics 24,* 1 (January), 98–117.

Tolani, D., Goswami, A., and Badler, N. I. 2000. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical Models 62,* 5, 353–388.

Vilhalmsson, H., Cantelmo, N., Cassell, J., Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thorisson, K., van Welbergen, H., and van der Werf, R. 2007. The behavior markup language: Recent developments and challenges. In *IVA 2007*.

Vogler, C. and Metaxas, D. 2004. Handshapes and movements: Multiple-channel american sign language recognition. In *Gesture-Based Communication in Human-Computer Interaction*. Lecture Notes in Computer Science, vol. 2915. Springer, Berlin, Heidelberg, 431–432.

Wang, J. and Bodenheimer, B. 2008. Synthesis and evaluation of linear motion transitions. *ACM Transactions on Graphics 27,* 1 (March), 1–15.

Wang, J., Drucker, S. M., Agrawala, M., and Cohen, M. F. 2006. The cartoon animation filter. *ACM Transactions on Graphics 25*, 1169–1173.

Warabi, T. 1977. The reaction time of eye-head coordination in man. *Neuroscience Letters 6,* 1 (October), 47–51.

Received December 2010;