Domain Specific Sign Language Animation for Virtual Characters

Anonymous

Keywords: Animation, Sign Language, Avatar, News Generation, Automatic

Abstract: This paper describes the design, implementation and results of a system to animate a virtual character to sign in International Sign (IS) By automatically parsing the input data, the system blends animations smoothly together in order to create a coherent and understandable presentation in sign language, with equal importance assigned to both hand and facial animation. Following the blending step, a video is rendered in a variety of formats suitable for distribution. The system was created in collaboration with groups of people with impaired hearing who were fluent in IS, and who were able to validate the results. We present a test case of the system in the shape of 'Borja', a virtual character who signed biweekly updates of the Barcelona World Race sailing regatta 2010/2011.

1 INTRODUCTION

The internet has transformed the way we access news. Be it a sports event, a major news story, or a conference or trade-show presentation, there are frequently several internet portals dedicated to providing live and interactive reports at a moment's notice. Indeed, this trend was visible as far back as 2004, where a survey of American internet users indicated that 53% checked sport news daily, and 55% of them did so through internet (Fallows 2004). This increase in news consumption has driven an equal rise in the number of services (Miesenberger et al. 2010)(Othman et al. 2010) that aim at creating news (and sports news) automatically from numerical data; generating stories which appear attractive to readers, while requiring minimal manual effort to maintain and update. Furthermore, this approach is particularly appealing for coverage of minority sports events, which typically do not receive the attention of more mainstream sports.

One disadvantage of this increased focus on automated news is that there is frequently no effort to make the news accessible for those users requiring more specialised services, such as those from the deaf community. According to the World Federation of Deaf, only 20% of the 70 million Deaf people in the world receive an education in spoken language(World Federation for the Deaf 2013), so the remaining 80% is not able to easily read and understand text written in spoken language (which is quite different from written sign language). While some effort is occasionally made on television to show expert signers providing translation of the news for deaf viewers, the limited resources available mean that this is rarely, if ever, duplicated across to internet-based media outlets.

In this paper, we present a virtual signing avatar designed to update users on the status of live events in a specific domain, such as sports events. The initial data sources are XML-based, featuring information such as the current score, competitor position in the event, or current prevailing conditions such as the weather. These data are sorted to fill a series of templates, which are input into our animation system. The system then automatically creates a video clip, with the data translated into animated sign language, and structured into logical sentences and phrases. Although the application of the system is domain-restricted, our work forms an important contribution to the on-going challenge of providing a fully automatic text-to-sign translation service.

Our system was tested live for a major sporting event, the Barcelona World Race sailing regatta 2010/2011. For the regatta, we designed an avatar and created a series of animation templates in IS. A fully automatic system parsed data from the race and, twice a week, renderered a video which was displayed on the website of the Spanish National

broadcasting company, Radio Televisión Española (RTVE). Results showed the pages featuring the animated clips to be the most popular in the Section focusing on coverage of the Barcelona World Race.

Our paper is organised as follows. Section 2 presents related work in the field of automated signing avatars. Section 3 details the methodology used to create our system. Section 4 presents Borja, the signing avatar designed for the Barcelona World Race. Finally, Section 5 summarises the paper and draws conclusions.

2 RELATED WORK

Several studies (Dehn & Van Mulken 2000; Johnson & Rickel 1997; Moundridou & Virvou 2002) found that rendering agents with lifelike features, such as facial expressions, deictic gestures and body movements may rise the so called persona effect. A persona effect is a result of anthropomorphism derived from believing that the agent is real and authentic (Van Mulken et al. 1998; Baylor & Ebbers 2003).

During the last decade there has been extensive research and development carried out with the goal of automating the animation of sign language for virtual characters. The ViSiCast and eSIGN projects (Elliott et al. 2007) focused on the development of a comprehensive pipeline for text-to-sign translation. Using text written in English and German as input, the system translates it into written text in sign language (English - ESL or German - DGS, respectively). The translation, though, is very literal and liable to produce unnatural results. Written sign language is translated into signs using the HamNoSys (Hanke 2004) notation which describes signs as positions and movements of both manual (hands) and non-manual (upper-torso, head and face) parts of the body. This notation enables signs to be performed by a virtual character procedurally, using inverse kinematic techniques. In this sense the animation system is procedural and, thus, flexible and reusable compared to others that use motion captured data or handcrafted animation. The HamNoSys notation, however does not include any reference to speed and timing, and ignores prosody. This has been considered one of the reasons of low comprehensibility (71%) of signed sentences using HamNoSys (Kennaway et al. 2007). This is a particularly serious disadvantage given that recent studies have demonstrated that prosody is as important in spoken languages as in signed ones, activating brain regions in similar ways (Newman et al. 2010), and has a crucial role in understanding the syntax of a signed message.

Automatic Program Generation is a field of natural interest to the commercial domain, but which has seen little academic research (Abadia et al. 2009). The most recent integrated attempt to disseminate sport news using a virtual signer has been the SportSign platform (Othman et al. 2010). It is a partially automatic system that needs an operator to choose the kind of sport and specify other relevant data, such as the teams playing match, the number of goals or points scored etc. Then the system generates a written version of the news in sign language (specifically, American ASL) that a human operator has to validate. Finally, a server-based service generates a video with an animated character that is then published on a web page. The workflow needs extensive interaction by a human user, in order to guide and validate the results of the system, meaning that it would not meet the important goal of reducing the costs to make signed sport news feasible.

3 METHODOLOGY

3.1 System Overview

Our animation system is designed to build a complex signed animation with a virtual character by concatenating, blending and merging animated clips previously prepared to digitally mimic the gestures of expert signers. The system relies on a series of XML-based templates indicating which signs should be used to build certain content, and which type of data is needed for the relevant information (numbers, text, etc.). Figure 1 shows a schematic overview of how the system constructs an animation.

The parser receives the data and decides whether is a known phrase, a number or a custom word. If it is a known phrase, it is retrieved directly from a database that stores the animation data for that entire phrase. If it is not a phrase, the system falls back and looks for a number or word. In this case, the sign elements that are used to compose that number or word them are taken individually and merged together to build a single new sign. This fall back is very useful when dealing with names of people or cities, or even with numbers of more than one digit. Finally, each of the clips is merged using animation blending. How each clip has to be blended is specified in the metadata description associated with that clip in the database. Each clip has its own length, start time, end time and in-out trim points which specify the boundaries for the blending.

3.2 Lexical Structure

Sign languages are complete languages with their own grammar and vocabulary. Communication occurs via a complex interrelation between manual, body and face gestures. These three elements cooperate as a whole to provide terms, grammar structure and intention (through prosody), and thus meaning. Even when written, sign languages are different from spoken languages. In fact, it can be difficult for pre-lingually deaf people to understand written English or other languages, as the grammatical structure of the spoken language is different to the structure of the sign language.

As many other languages, vocabulary and grammar also change depending of the country (ASL is different from Spanish Sign Language LSE) and it may also adapt to the characteristics of a certain zone inside a country (non coastal regions of a country may lack the term "boat"). In an attempt to provide a common standard of communication for the world deaf community, in 1975 the World Federation of Deaf proposed a unified system which included the most common terms used in the majority of different sign language. It also included signs that are easy to understand. Since then, International Sign has evolved and is now used in conferences around the world and for sport reviews (as in the 2010 Fifa World Cup).

In this sense, the lexicon of International Sign System is quite poor but its grammar is as reach and complex as for any other language (Newman et al. 2010).

3.3 Data Parser

The goal of any automatic news system is that it should be able to take data from automatically updated sources, and transform that data into a manner such that it can be broadcast and easily understood by users. Our system parses XML data with a specific format (see Figure 1 below) and uses it to fill in a series of data templates. These templates are structured in a similar way to the lexical structure of IS as mentioned above. The data is now in suitable format to be converted to animation.

3.3 Animation Database

There are three possible sources for the animation data suitable for our system:

1. *Motion Capture Systems*, where an actor's performance is digitally captured
2. *Procedural Systems*: where the actual animation is generated by a fully automatic system
3. *Manual recreation from video*, where a human animator manually creates animations using a video of an actor as a reference.

Standard Motion capture systems do not provide detailed information on the movements of the fingers and of the facial muscles: both essential for accurate representation of sign language. Sign language makes use of incredibly subtle movements

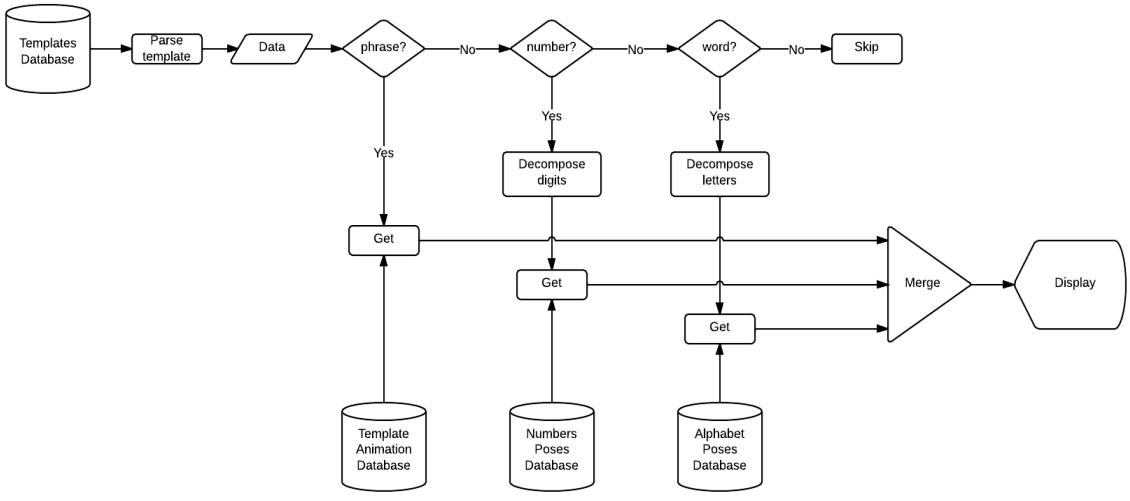

Figure 1. The animation system receives an XML file that contains the description of the clips that has to be concatenated in order to build the complete animation. Depending on whether it's a pre-made phrase, a number or a word, the system retrieves and merges the appropriate animated clips to build the signed contents.

of the arms, fingers and facial muscles, which requires considerable effort to capture accurately. Although full performance capture technology is being developed (Weise et al. 2011), combining facial and finger systems to the required standard is beyond the scope of this work. Furthermore, raw motion capture data contains many small errors and rarely can it be used in its native form. The effort required to manually 'clean' the capture data, adapting it to a virtual character and enhancing it with the missing pieces (commonly face and fingers) is a highly time-consuming process. Thus, we discarded motion capture as an animation source

Procedural animation has also been recently used for automated sign language generation (Elliott et al. 2007). It is a flexible solution as it does not require an expert to sign the contents, and any sign can be generated using a sign notation like HamNoSys (Hanke 2004). Unfortunately, as mentioned above, animations generated with sign notations suffer from a notable lack of prosody, a very important factor in sign languages.

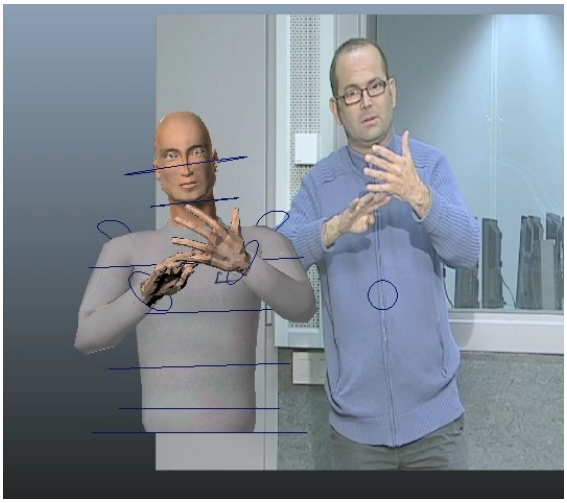

Figure 2. Screenshot of a signing expert (prelingually deaf) with virtual character overlay.

This means that every animation clip used by the system has to be animated by hand by following video references of signing experts recorded on video. We created a skeletal animation rig for the character body which uses both forward and inverse kinetmatics. We use blend shape animation for the the facial expression as this gives us greater control on the precise shape of the face – an essential component for the understanding of sign language. While this requires initial manual effort, the main motivation behind the choice of this technique is the freedom and accuracy that hand-made animation can provide. Furthermore, as the 'library' of animations is created, adding to it becomes progressively easier, as many previous animations can be reused. Figure 2 shows a screen capture of the animation process. It is important to note that, from a technical point of view, our system is capable of working with both manually created animations, motion-captured performances, and procedural animation.

For animation purposes, we define four different semantic levels:

1. *Phrases*: where the actor signs a complete or partial phrase e.g. "My name is…"
2. *Words:* where the actor signs an entire word e.g. "weather"
3. *Letters*: individual letters of the alphabet
4. *Numbers*: single digit numbers only.

We distinguish between words and phrases because, frequently, sign language will employ a single sign to encapsulate an entire phrase of spoken language. The advantage of these distinctions is that they are designed to work very effectively with the 'trickle down' animation composing/blending. Animation clips are designed to unequivocally communicate a single chunk of information, which can be understood in isolation, which is essential for composition and blending (see below).

Once recorded and animated, each clip is stored in an Animation Database along with its associated metadata (mainly the spoken language equivalent of the animation).

3.4 Composing and Blending

To compose a sign language clip from spoken language input, our system is split into three components (the Classifier, Blending Queue and Composer) which process any input ready to be passed to the renderer (see Figure 3). The Classifier first analyses the input template according to the four semantic levels defined above. From this classification, the system searches the Animation Database in a trickle down manner, (as shown in Figure 1 above). Let us take the example "My name is John". The classifier first parses the entire template to search for a known phrase in the database. It finds the phrase "My name is" and adds it to the Blending Queue, but does not find the phrase "John". It then drops down a semantic level to look for any numbers. In this case, there are none, so it drops down another level to search for the word "John". If it does not find the word, then it drops down a final level and adds the individual letters 'J', 'O', 'H' and 'N' to the Blending Queue.

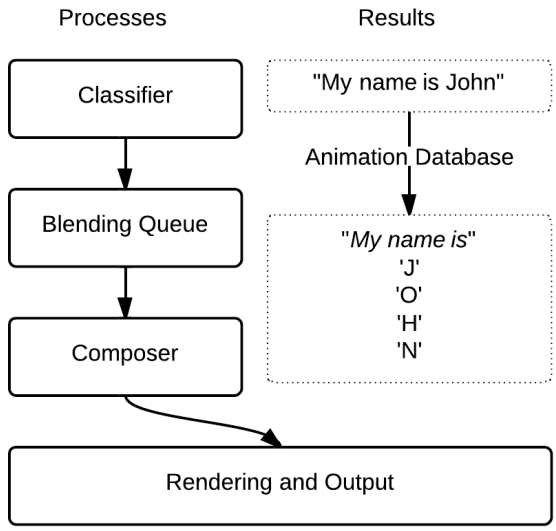

Figure 3. The flow of information through the different processes that create an animation clip from input

In order to concatenate each clip to conform a complete discourse, The Composer must blend the clips in the Blending Queue together to perform smooth transition between different animations. Blending can be performed in two different ways: *Additive blending*, where an animation clip is added on top of an underlying clip, offsetting it by means of differences between each other; and *Override blending*, which substitutes a given animation clip with the content of another one. Furthermore, the blending can be performed either between two clips or across two clips (see Figure 4). When working with sign language, it is important to consider that even minor modification of the sign could change the meaning for the viewer (potentially drastically). To avoid this, we decided to concatenate phrases performing override blending between clips. In this way we guarantee that the signed content is 100% accurate and that no undesired signs are produced in the transition between one clip and the other. On the other hand, words and numbers are built by blending together multiple short clips of letters and digits. In this case, we blended across clips, which produced smooth and continuous (not robotic) signs

Figure 4. Figure shows how blending can be made between two clips. Left: interpolating from last frame of Clip A to first frame of Clip B; or Right: cross fading from Clip A to Clip B.

3.5 Rendering Output

The system is designed to render a video output using standard algorithms from Autodesk Maya. Background can be static or animated in a pre-rendering step. This approach provides acceptable quality results with a relatively low rendering times, which is crucial for providing large amount of content in a reasonable time.

4 BORJA: SPORTS REPORTER

Our system was tested in a real environment for the Barcelona World Race sailing regatta, 2010/2011. The Barcelona World Race is a non-stop, round-the-world regatta, where crews of two people circumnavigate the globe in an IMOCA 60 racing boat, competing to be the first to return to Barcelona. Apart from being a challenging regatta, the event also has at its heart a commitment to research and development, education, and science and the environment. In that sense it was an ideal opportunity for us to test our system in a real-world context. The race was forecast to last 3 months, and consisted of 14 boats. Our goals, therefore, were to:

- Design and animate a virtual character suitable for presenting news from the regatta, in sign language;
- Adapt our system to be used for automatic creation of news from a sailing regatta;
- Create an automatic, server-based system that would output a video of the current animated news, twice a week, and make it available for public dissemination;

4.1 Character Design

Any virtual character has to be designed in a way that makes it believable and, for a virtual newsreader, in a way that helps convey information to the audience. The character design should match the characteristics of the sport and adapt to different context, for instance changing clothing, depending on external factors such as the weather.

Another aspect, particularly important for virtual signers, is the believability of the movements and gestures of the character. It has to look convincing (and not necessary *realistic*), in order to encourage the audience to engage with the provided news. This insistence on believability was a major factor in our deciding to create personalized hand-crafted

animations for the character, rather than procedurally animated signs that lack prosody and life.

Another crucial aspect is the size and proportions of the hands. The automatic contents generated with our method were going to be viewed mainly through a video streamed on a web page, meaning that the playback area would not be large enough to ensure that the hands and fingers will be clearly visible all the time. On the other hand, larger hands assist in driving attention toward such an important communicative item. Thus, we ran several tests with experienced signers in order to decide on the correct proportions for all aspects of the body. After several iterations, we settled on a blond male, dressed in typical sailor garb (varying depending on current weather), and christened him *Borja*.

Figure 5. Art design for Borja. Design includes different clothing for different weather conditions. The design shows a sporty style for the character to make it closer to the sailing world, bigger eyes to empathize with audience and big hands to make signs more readable through streaming videos.

4.2 Input Data and Template Creation

There were three main data sources used for this application, all provided under license by the race organisers:

- Boat XML information, with direct information about each competitor boat and race info per boat, such as GPS location, speed, course, ranking place, distance covered etc;
- The GRIB weather data, updated every 30 minutes;
- Tertiary information regarding the race e.g. historical information regarding previous races, skipper biographies etc.;

The system groups the data needed for each news and ranks the 'relevance' of each set of data. The relevance is calculated according to criteria pre-set by expert news writers focusing on sailing regattas. For example, a change of leader is considered highly relevant, while a change at the bottom of the ranking is less relevant. Once these criteria are set, no further expert input is required, and the system generates rankings automatically. Based on the calculated ranking, the system picks the three most relevant items of news, and adds *opening* and *closure* sections to build the complete template for animation (see Figure 6). In the case where the third less relevant news is below a relevance threshold, the system provides information about the skipper, the boat or the geography.

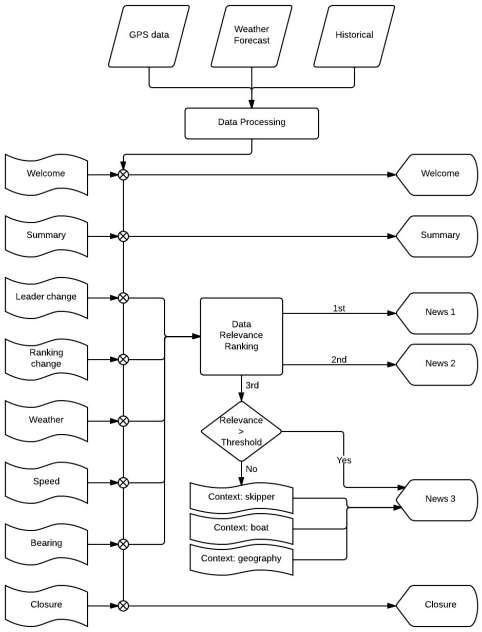

Figure 6. Scheme of the news generation system.

4.1 Scheduled Output

During the course of the regatta, the system compiled the data twice a week, every week during a period of three months (the length of the regatta). From this data, the system reads the prominent weather condition and configures Borja to wear different clothes (wet jacket, t-shirt, etc.) depending on it. Then, the system concatenates a sequence of animations to represent the desired news in IS and launches the rendering. All the information about how to build the animation sequences and the database of available animations is stored in XML files.

Once the rendering is completed, the video (in Adobe Flash *.flv* format) and the equivalent written news in Spanish is sent to the server of RTVE to be

displayed on their webpage (see Figure 7). Figure 8 shows a larger screenshot of Borja in action.

Figure 7. Webpage of the Spanish National Radio showing *Borja*'s videos for the Barcelona World Race.

Figure 8. Screenshot of Borja in action

5 RESULTS AND CONCLUSIONS

In this paper, we present the design and implementation of a system capable of automatically transforming input data into sign language, given a specific, restricted domain. The system is designed to automatically create reports for live events such as sports events, and we present a successfully prototype, implemented to create automatic signed news for the Barcelona World Race 2010/2011.

Figure 9. Screenshot of RTVE ranking page showing Borja popularity.

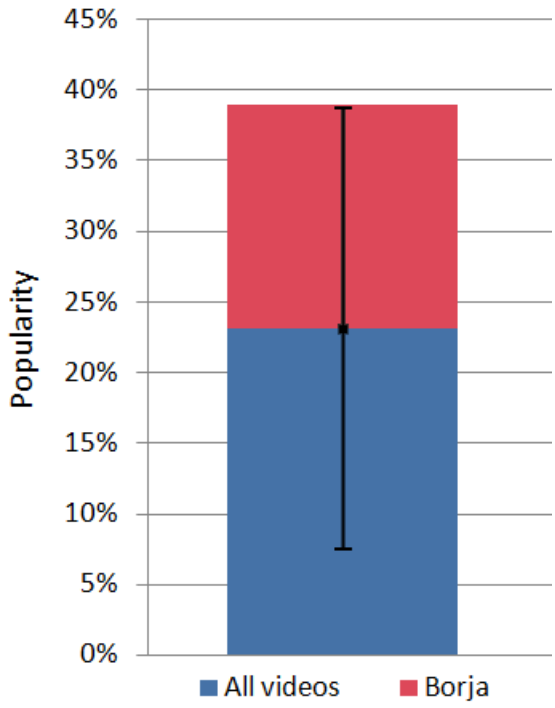

Figure 10. Chart showing the 15.8% greater in popularity of Borja's videos over all other Barcelona World Race multimedia. Also shown in the Standard Deviation bar.

25 videos featuring Borja were created during the Barcelona World Race and featured on the RTVE website. These 25 formed part of a total of 91

multimedia entries concerning the Barcelona World Race 2010/2011. The 'popularity' of the multimedia entries is habitually measured by RTVE by measuring web traffic on each page. The results show that the mean popularity rating for the 25 Borja videos was 38.96%, 15.8% greater than the mean popularity of all 91 multimedia videos. This increased in popularity is marginally greater than the dataset standard deviation of 15.6%.

The output of the system was validated as "consistently correct" by the expert signers involved in the project, but future work is now focused on obtaining more quantitative results as the quality of the animated output. A study featuring a statistically valid sample of deaf users is being carried out in order to ascertain statistically whether the animated sign language produced by the system is capable of being understood naturally in different countries around the world. If the results of this analysis are positive, we plan to develop our system and use for different scenarios such as rolling news and other live events.

ACKNOWLEDGMENTS

[*Space reserved in this section for acknowledgments in final paper*]

REFERENCES

Abadia, J. et al., 2009. Assisted animated production creation and programme generation. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology ACE 09*. ACM Press, p. 207.

Baylor, A. & Ebbers, S., 2003. The Pedagogical Agent Split-Persona Effect: When Two Agents are Better than One. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*. pp. 459–462.

Dehn, D.M. & Van Mulken, S., 2000. The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52(1), pp.1–22.

Elliott, R. et al., 2007. Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4), pp.375–391.

Fallows, D., 2004. The Internet and Daily Life. *Pew Research Center's Internet & American Life Project*. Available at: http://www.pewinternet.org/Reports/2004/The-Internet-and-Daily-Life.aspx. [Accessed July 24, 2013]

Hanke, T., 2004. HamNoSys - representing sign language data in language resources and language processing contexts. In *LREC 2004, Workshop proceedings : Representation and processing of sign languages*. pp. 1–6.

Johnson, W.L. & Rickel, J., 1997. Steve: an animated pedagogical agent for procedural training in virtual environments. *ACM SIGART Bulletin*, 8(1-4), pp.16–21.

Kennaway, J.R., Glauert, J.R.W. & Zwitserlood, I., 2007. Providing signed content on the Internet by synthesized animation. *ACM Transactions on Computer-Human Interaction*, 14(3), p.15–es.

Miesenberger, K. et al. eds., 2010. *Computers Helping People with Special Needs*, Berlin, Heidelberg: Springer Berlin Heidelberg.

Moundridou, M. & Virvou, M., 2002. Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, 18(3), pp.253–261.

Van Mulken, S., André, E. & Müller, J., 1998. The Persona Effect: How Substantial Is It? *People and Computers*, XIII, pp.53–66.

Newman, A.J. et al., 2010. Prosodic and narrative processing in American Sign Language: an fMRI study. *NeuroImage*, 52(2), pp.669–76.

Othman, A., El Ghoul, O. & Jemni, M., 2010. SportSign: A Service to Make Sports News Accessible to Deaf Persons in Sign Languages. *Lecture Notes in Computer Science*, 6180, pp.169–176.

Weise, T. et al., 2011. Realtime performance-based facial animation. In *ACM SIGGRAPH 2011 papers on - SIGGRAPH '11*. New York, New York, USA: ACM Press, p. 1.

Williams, R., 1957. *The Animator's Survival Kit*, Faber and Faber Inc.

World Federation for the Deaf, 2013. Human Rights. Available at: http://wfdeaf.org/human-rights [Accessed July 24, 2013].