SIGNSPEAK Project Tools: A way to improve the communication bridge between signer and hearing communities

Javier Caminero[1], Mari Carmen Rodríguez-Gancedo[1], Álvaro Hernández-Trapote[2], Beatriz López-Mencía[2]

[1] Telefónica R&D, Madrid, Spain

[2] Universidad Politécnica de Madrid, Spain

Email: fjcg@tid.es, mcrg@tid.es, alvaro@gaps.ssr.upm.es, beatriz@gaps.ssr.upm.es

Abstract

The SIGNSPEAK project is aimed at developing a novel scientific approach for improving the communication between signer and hearing communities. In this way, SIGNSPEAK technology captures the video information from the signer and converts it into text. To do that, SIGNSPEAK consortium has devoted great efforts to the creation and annotation of the RWTH-Phoenix corpus. Based on it, a multimodal processing of the captured video is carried out and the resultant sign sequence is translated into natural language. Afterwards, the intended message could be communicated to hearing-able people using a text-to-speech (TTS) engine. In the reverse way, speech from hearing-able people would be transformed into text using Automatic Speech Recognition (ASR) and then the text would be processed by virtual avatars able to compose the suitable sign sequence. In SIGNSPEAK project, scientific and usability approaches have been combined to go beyond the state-of-the-art and contributing to suppress barriers between signer and hearing communities. In this work, a special stress was put in the development of a prototype and also, in setting of the grounds for future real industrial applications.

Keywords: tool development, sign-language-to-text, user evaluation

1. Introduction

Communication for Deaf community is based on sign language since it is "the only language Deaf people can acquire effortlessly and spontaneously when given the right input" (Wheatley and Pabsch, 2010). Unfortunately, deaf and hard of hearing signers have serious limitations for communicating with people without no sign-language skills and thus, the integration into educational, social and work environments is not complete.

Although the mother tongue is defined as the first language that one has acquired, for the deaf community, it is more complex than that, then only a small percentage of deaf children acquire a sign language naturally and in similar stages as hearing children do with a spoken language.

Taking into account these peculiarities, we realize that deaf people usually find numerous barriers in communication. Some of these barriers include the presence of an operator (which may be seen as intrusive and do not represent parity with hearing people), slow communications connections and lack of awareness of how to communicate with people who are deaf or have speech difficulties. About these issues, recent studies (Market Research, 2011) reveal that deaf people and individuals with speech difficulties need freely-accessible services and equipment to ensure that their communication needs are totally fulfilled.

Having all these issues into consideration, SIGNSPEAK project[1] is aimed to provide deaf people a communication bridge between signers and hearing communities. Thus, a new vision-based technology for translating continuous sign language into text is being developed. For that purpose, it has been needed the creation of RWTH-Phoenix, a suitable video corpus for data-driven automatic sign language processing (Stein et. al., 2010). As a consequence of the automation of the services and applications provided by the SIGNSPEAK technology, users' privacy feeling and their confidentiality in the communication process would be improved.

2. SIGNSPEAK: establishing a new communication bridge

As it is showed in Figure 1, SIGNSPEAK technology captures the video information from the signer and converts it into text. In order to do that, a multimodal processing of the video is carried out and afterwards, the resultant sequence of signs is translated into natural language. Using a text-to-speech (TTS) engine, the intended message is communicated to people who are able to hear. In the reverse way the speech from hearing community is captured and translated into text (Automatic Speech Recognition-ASR). Then the text is used by virtual signers (avatars) which compose the suitable sequence of signs.

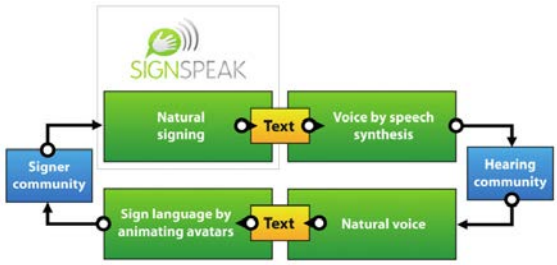

Figure 1. Communication bridge between Signer and Hearing communities

[1] http://www.signspeak.eu

2.1 Text-to-Speech

A text-to-speech (TTS) synthesiser can be defined as a piece of software which transforms into speech any input sentence in text format (Dutoit, 1997). This functionality makes a TTS very useful for communication systems because it avoids pre-recording every sentence or words planned to be used in a service. There is a wide availability of products, i.e. Loquendo TTS, Nuance Vocalizer or Festival.

Despite of the great performance of the aforementioned systems, there are yet critics to the use of TTS for certain applications due to: pronunciation of new and rare words (Spiegel, 2003), prosody (Hirschberg, 2002) or limited availability in certain new languages.

2.2 Automatic Speech Recognition

The human voice is generated by the vibration of the vocal cords. The vibration of the cords moves the air and these variations of pressure arrive to the listener's ear. Then the pressure waves are transformed into a signal that is processed by the brain and properly interpreted. The acoustic features of this signal allow the listener to differentiate one sound from another, and that is what an Automatic Speech Recogniser (ASR) tries to accomplish.

Some of the most relevant actors in the development of this technology are: CMU Sphinx, RWTH ASR, Dragon Naturally Speaking or Microsoft Speech API.

However, the performance of an ASR system usually depends drastically on external factors (Acero, 1992): input level, additive background noise, channel distortion, etc.

2.3 Signing Avatars

Recently, the virtualization of everyday life and the gaming industry has promoted a great development of the virtual characters field. The improvement of several communication technologies as the automatic speech recognition or the text-to-speech engines makes it real to create virtual agents able to interact with users. The benefits are obvious: cheaper customer service and 24/7 availability. Furthermore, through this kind of interfaces, users could establish relationships close to those ones between humans (Reeves and Nass 1996).

Some applications which use avatars or a sort of them for representing Sign Language are Sign Smith Studio, Sys Consulting, ViSiCAST, eSIGN, DePaul ASL Project, SignSynth, TEAM, etc.

Figure 2. ViSiCAST signing avatar

2.4 Sign Language to Text

The most challenging technology included in the communication bridge proposed by SIGNSPEAK is the translation of Sign Language into text. That is, capture the movements, expressions and emotions of the signers; identify the signs from the extracted features, and then translate the sequence of them into natural language in order to obtain a message understandable by hearing users.

The means used to capture hand movements can be classified mainly in two groups: instrumented and video-based. For instrumented proposals, gloves are usually complemented with other devices, as accelerometers. It means users have to remain close to the radiant source, in the case of a wireless connection, or close and physically tethered to the computer in the case of a wired one. Furthermore, current glove technology is not intended for daily use; the gloves deteriorate quickly with extended use and output becomes increasingly noisy as they break down. In the other hand, this kind of solution uses to be more reliable, overall against ambient noise or other adverse background conditions. In video-based approaches, the signer avoids having attached to hers/his body any instrumentation. However, the working conditions should be controlled and the amount of data obtained, compared with instrumented systems, is lower.

SIGNSPEAK project wants to go beyond most of the limitations previously presented. The project follows a global planning approach to transfer the technology to the daily life of deaf community and its scope implies advances in several research fields and the need of taking into account the industrial perspective.

3. Main Technological Factors

The SIGNSPEAK project is intended to be a first step to achieve a sophisticated technology able to complete the communication bridge between hearing and deaf community. In this preliminary stage, the demands about the performance of the technology should be ambitious but bearing in the mind the possible problems which could arise in a realistic scenario. Thus, for a proper operation of the technologies involved in the communication bridge (SIGNSPEAK, TTS, ASR, avatars) different user and environmental factors and some technological limitations need to be considered. Next, we point out some of them, however for more detailed information refer to (Gancedo, Caminero and Van Kampen, 2011).

3.1 User factors

User factors are individual differences that include demographic variables and situational variables that account for differences attributable to circumstances such as experience and training (Agarwal and Prasad, 1998).

Some user factors which could be relevant to SIGNSPEAK technology might be:

- *Gender*. Research has shown that there are differences between men and women regarding the cognitive structures employed during the interaction with

technology products (Venkatesh and Morris, 2000). For example regarding signing avatars, in (Bailenson and Yee, 2005) is showed how users prefer avatars which are similar to them and mimic their behavior. Thus, it could be suggested that signing avatars should mimic the signing style of users or even adopt the users' gender.

- *Experience with technology*. For automatic speech recognition (ASR), in (Karat et al., 2000) is described an experiment where the ASR performance is worst for novice users than for expert ones. Furthermore, the latter group of users is more effective carrying out the corrections when the system fails.
- *Age*. From the point of view of acceptance of technology, age is recognized as a key factor. Specially, senior users, who do not usually have great experience with technology and have age-related problems with cognitive abilities, face difficulties understanding and interacting with technological devices (Ziefle and Bay, 2008). On the contrary, older users are more inclined to accept technologies when the usefulness is clear and there is a good support of the system (tutorials, help system, etc.) (Arning and Ziefle, 2010).
- *Cultural background.* For ASR engines, the problematic issue is quite clear: the accent. In (Huang et al., 2001) the accent was identified as one of the principal components of speech variation.
- *Other factors*. Many more user factors could affect the acceptance of a new communication paradigm (i.e. SIGNSPEAK's communication bridge). For example the level of signers' expressiveness, the users' emotional state or the users' physiology.

3.2 Environmental factors

The conditions relative to the context where the interaction is performed are collectively called environmental factors. They include numerous variables as weather conditions (i.e. lighting), noise conditions (i.e. "the cocktail party effect") or location conditions (i.e. mobility, in-car scenario…).

In the case of TTS and ASR engines, arguably the most harmful effect is that posed by noisy environments. Regarding virtual signers, taking into account that deaf users should be looking with attention to the virtual agent, the cognitive load that the environment demands has to be taken into consideration. In order to illustrate this, let's imagine an application designed for interacting through a tactile interface and that uses at the same time a virtual signer for communicating the information. Then, it is necessary to set the message of the virtual signer in such a way that it does not coincide with any other visual message. For SIGNSPEAK, there are three main tasks related to the multimodal visual analysis: tracking of hand positions, facial analysis and body pose estimation. All of them need robust tracking algorithms, since they should avoid the effect of i.e. signing hands moving in front of the face, or signing hands crossing the other hands.

3.3 Resource-related factors

Among these factors, we can include the computational power or memory availability in the devices or quality of communication requirements. Thus, these resources have influence on the selection of a concrete technology, the use of a concrete device (i.e. a desktop environment vs. a portable device) or in a worst case scenario, a degradation of the user perceived quality.

In the case of SIGNSPEAK technology, very demanding requisites regarding computational power are needed. Its flow network implies several stages with certain complexity. Due to that, there is a delay factor of around 20 times compared to real-time (for example, the translation of 6 seconds of video will take around 2 minutes) for testing data coming from the same domain as the data used to train the system.

3.4 User perception and acceptance of the technology

User acceptance of a new technology does not depend exclusively on its technical functionality. User perception of a new technology is built from a set of psychological, social and contextual factors that are related to its use in everyday life applications. Some of these factors have been already mentioned in section 3.1 although complete models from different perspectives and at various levels have been developed (Venkatesh et. al., 2003).

Results of an expert survey performed in SIGNSPEAK project, regarding what aspects are more important for selecting technological products are showed in Figure 3. These results are presented through a bar graph showing three colours depending on the relevance for the user (red for low relevance, orange for mid-range relevance values and green for high relevance).

After analysing the results, we can see that the price of a product is not seen as a fundamental factor and apparently when the service provided by the technology is really useful, price is not very important. Logically this fact, that is applicable to any target population, gains importance for the deaf community since technology helps them break down very annoying communication barriers. At the other end, the more relevant factors are: usefulness, easiness of use and having the ultimate technology. Related with the abovementioned great need of technology products able to help the deaf community, it seems clear that usefulness is a key variable for choosing a device. This may indicate that cutting-edge technology has been associated to a perceived loss of reliability of the technology's performance (a factor which was not represented in the survey). And finally, due to the accessibility difficulties which traditionally deaf people have to face in the use of technology products, easiness of use is also lightly highlighted.

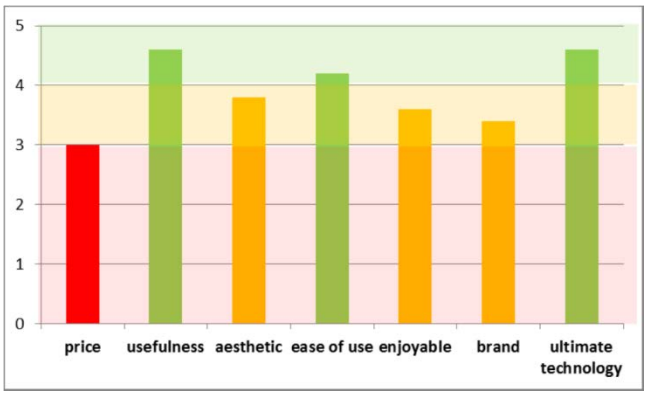

Figure 3. Bar graph about the most relevant factors in the selection of a technological product.

4. Application Scenarios

In order to select a relevant application scenario for applying SIGNSPEAK technology, an expert survey was performed. The creation of the questionnaire and the selection of the experts were made jointly between European Union of the Deaf (EUD) and Telefónica R&D (TID), both partners of SIGNSPEAK project.

In section 4.1 a review of the communication problems highlighted by these experts is presented. Later, in section 4.2, the experts' feelings about a set of possible application scenarios are listed.

4.1 Communication problems

The expert survey addresses the issue about the more unpleasant situations when deaf people have to communicate with non-signers.

Some of these situations are listed below:

- Telephoning hearing people through a relay service based on text, which is not their first language.
- Sending e-mails via text messages, instead of using their first language.
- Video films using sign language are often not subtitled and most of the hearing people cannot understand sign language.
- Accessing to public authorities/services (i.e. passport issuing service, banking, etc.) where most people cannot sign.
- Relay services almost never opens 24 hours a day.
- Hearing people cannot learn sign language without getting instructions in their own written or spoken language.

4.2 Scenario analysis

In the expert survey several scenarios were proposed. These scenarios, created in collaboration with EUD, take into account the communication needs of deaf community and the forecasted usefulness.

All of them have the following motivation story:

"John and Mary are a deaf-hearing marriage and they have one child, Susan, who is 7 years old and she is also deaf. This family is bilingual; sign language and spoken language. They have hearing neighbours and family members who cannot sign very well."

4.2.1 Sign language e-learning

This scenario is as follows:

"A neighbor girl of Susan is following a course for improving her sign language skills. For doing this course, pupils have to connect to the teacher through Internet (using a webcam). Then, pupils see the teacher in their monitors and the teacher can see all the pupils at their own homes. The teacher gives the lessons using sign-language and, thanks to SIGNSPEAK technology, text subtitles appear at the same time."

In this case, experts told us *"beginners learn better with signing videos without subtitles and then they can watch signing videos with the subtitles to see if they already understand sign language"*.

4.2.2 Answering machine

This scenario is as follows:

"John is in a congress and makes a video call to home. Nobody is at home, so he leaves a recorded video with his sign language message. The answering machine, through SIGNSPEAK technology, translates the sign language message into text. When Mary arrives home, she realizes there are several messages. As she is busy, she decides listen the messages while preparing the dinner. She listens to her husband's message through a voice synthesizer."

This service arouses a similar feeling as Sign Language e-learning, at least for those who can sign very well, since *"I would prefer to see him directly signing instead of hearing the voice synthesizer"* betting for the concept of a more realistic conversation. Therefore, for someone not able to sign well or at all, a service like this would be considered as a good idea

4.2.3 Play Sign Language

This scenario is as follows:

"Susan has a game console which includes a camera. She wants to play with her neighbour girl. They love to play an educative adventure game that makes you practice some sign language expressions. Using the video from the camera, SIGNSPEAK technology assesses the quality/correctness of the signs and the game gives Susan feedback about how to improve her sign language abilities. As the neighbour girl gets better, she moves forward the levels of the game. They improve their communication very well through playing the game."

Finally the game for practicing Sign Language was really welcomed since *"playing with sign language is the best way to learn it. If it is more formal as in the school, then children would get very bored."*

4.2.4 VideoSL mail

This scenario is as follows:

"Mary wants to send an email to several people. Some of them can hear while others cannot. She records a video signing and she sends it. SIGNSPEAK technology

translates the sign language message into text and then it sends the email with the video and the text message to all the addressees."

VideoSL mail was considered as good and suitable to SIGNSPEAK technology. Additionally, it was detected as a possible application for learning: *"Hearing people would learn sign language by reading the text. Text and sign language should be next to each other in the system."*

5. The Prototype: VideoSL Mail

Since one of the main goals of SIGNSPEAK project was to analyse the industrial application of SIGNSPEAK technology in order to fully understand the possible implications of the integration of this technology, finally, based on the experts' opinions and the limitations of SIGNSPEAK technology (i.e. non-real-time processing), the VideoSL mail scenario was selected for the development of a prototype. This scenario was devised as employing a similar concept of use as Google Voice automatic voicemail transcription, helping a signer-hearing group of friends to socialise together without the need for interpreters.

The main advantages pursued with this prototype are:

- Text previewing of the video messages. This feature is particularly oriented to those deaf people comfortable with reading.
- Ability to search of information in video data.
- Allowing deaf people to express themselves using Sign Language.
- Making it possible that non-signers hearing people can understand a message expressed in Sign Language.

5.1 Architecture

Telefónica R&D has implemented a framework and user interface based upon many of the principles of cloud computing. This framework will provide a flexible communications infrastructure for developing SIGNSPEAK services. Cloud computation is defined as the provision of computing services over the Internet in a manner reminiscent of those of public commodities such as electricity or watering systems. Thus, it is a way to offload processing of data to places other than the user's system.

In Figure 4, the general architecture diagram devised for SIGNSPEAK service is shown. The devices communicate directly using a web-based interface such as a browser (in the case of a traditional PC) or using a mobile application that adapts the UIs to the particularities of a mobile device (that could be either a tablet or a smartphone). These applications communicate their translation requests (in the figure, this channel is marked with green arrows) via the web interface to the SIGNSPEAK servers. The web interface is a standard Web Service that accepts basic data such as the stream/location of the input video and some settings for the translation (e.g., addressee or timing constraints). The Job Scheduler is a module that gathers all the translation requests and generates a list of "translation jobs" to be executed. Finally, the jobs are sent to the SIGNSPEAK translation pipeline and the results are tored into the database, being available at request.

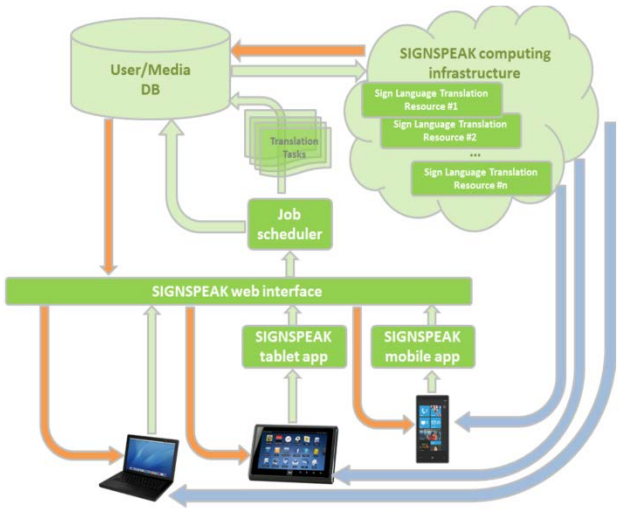

Figure 4. The SIGNSPEAK Cloud architecture

5.2 User interface

In order to simplify the portability of the interfaces between different devices and platforms, the user interface was implemented using HTML5 and JavaScript (making use of Sencha Touch library). Its main features are:

- Easy user interaction. Mails are presented in a vertical carousel, so users can use up/down swipe gestures to view the mails.
- Filtering capabilities. The search functionality is accessed through a text box and it makes possible the e-mail filtering based on the e-mail bodies or on the translations generated by SIGNSPEAK.
- Quick use feature. Users can select some videos as a sort of frequent replies and then attach them to their e-mails using a drag-and-drop paradigm.

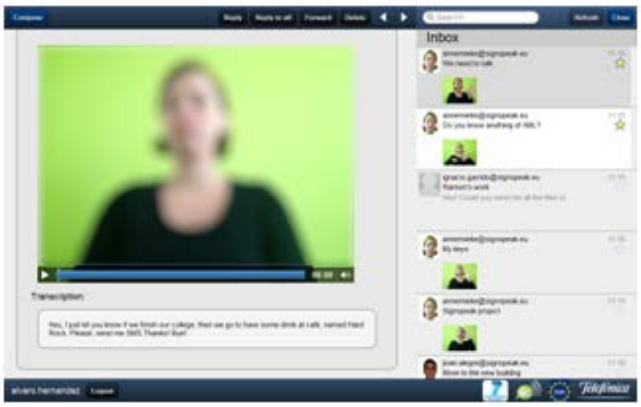

Figure 5. VideoSL mail user interface

5.3 User Experience evaluation

In collaboration with EUD a user evaluation has been carried out to gain insights about the suitable performing and acceptance of the VideoSL mail service.

The prototype was installed into a touch tablet device and a task-guided evaluation was carried out by 5 users. Once they had interacted with the application, they filled out a questionnaire. Some of the factors addressed by this preliminary evaluation are:

- Previous experience regarding email services and tablet PC devices.
- Likeability of the service.
- System performance.
- Usefulness of the service.
- Willingness to buy.
- Overall acceptance.

After gathering and interpreting users' feedback, the first results show a high acceptance and excitement about this system and how its daily life would be much easier thanks to the use of this technology.

6. Conclusions

In this paper, SIGNSPEAK project has been presented, focusing on one of its main challenges, i.e. how to improve the communication bridge between signer and hearing communities. Telefónica R&D as the main industrial partner of the project has addressed this challenge, firstly studying the main needs of potential users, and then creating an application prototype of a VideoSL email service, still without full functionality due to the limitations of the state-of-the-art technology for a real-time operation, but able to provide a similar User Experience to that than a real service would cause.

A preliminary user's feedback has been collected, showing how excited they are about this prototype, but also making us aware of the necessity of continuing the Research on this technology field.

7. Acknowledgements

This work received funding from the European Commission's Seventh Framework Program under grant agreement number 231424 (FP7-ICT-2007-3).

The authors express their gratitude to the European Union of the Deaf (EUD) for its cooperation in the framework of SIGNSPEAK project, providing first-hand useful information and access to real users and their needs.

8. References

Acero, A. (1992). Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers

Agarwal, R. and Prasad, J. (1998). A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology. Info. Sys. Research (INFORMS) 9: 204-215

Arning, K. and M. Ziefle (2010). Ask and you will receive: Training older adults to use a PDA in an active learning environment. International Journal of Mobile Human-Computer Interaction 2, no. 1: 21-47

Bailenson, J.N., and N. Yee (2005). Digital chameleons, Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. Psychological Science (SAGE Publications) 16, no. 10: 814

Dutoit, T (1997). An introduction to text-to-speech synthesis. Vol. 3. Springer

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. Speech Communication 36, no. 1-2: 31-43

Huang, C., Chen, T., Li, S., Chang, E. and Zhou, J. (2001). "Analysis of speaker variability." Citeseer. 1377-1380

Karat, J., Horn, D.B., Halverson, C.A. and Karat, C.M. (2000). Overcoming unusability: developing efficient strategies in speech recognition systems. ACM. 141-142

Market Research (2011). Opinion Leader. "OfCom Relay Services". UK competition and regulatory authority, Market Research

Reeves, B., and Nass, C. (1996). The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press New York, NY, USA

Rodríguez-Gancedo, M.C., Caminero, J., Van Kampen, A. (2011). Report about the study of the new communication bridge between signers and hearing Community. Signspeak Project, Deliverable D9.4. http://www.signspeak.eu/deliverables/SIGNSPEAK_D9.4_v1.0.pdf

Spiegel, M. F. (2003). "Proper Name Pronunciations for Speech Technology Applications." International Journal of Speech Technology (Springer Netherlands) 6: 419-427

Stein, D., Forster, J., Zelle, U., Dreuw, P. & Ney, H. (2010). RWTH-Phoenix: Analysis of the German Sign Language Corpus. 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Malta, May

Venkatesh, V. and Morris, M. (2000). Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. MIS Quarterly (Management Information Systems Research Center, University of Minnesota) 24, no. 1: pp. 115-139

Venkatesh, V., Morris, M., Davis, G. & Davis, F. (2003). User Acceptance of Information Technology: Toward a Unified View, MIS Quarterly 27(3), pp. 425-478

Wheatley, M. and Pabsch, A. (2010). "Sign Language Legislation in the European Union". Edited by EUD. European Union of the Deaf

Ziefle, M. and Bay, S. (2008). Transgenerational designs in mobile technology. Handbook of Research on User Interface Design and Evaluation for Mobile Technology 1: 122-141