SIGNSPEAK

Scientific understanding and vision-based technological development for continuous sign language recognition and translation

Grant Agreement Number 231424

Small or medium-scale focused research project (STREP)
FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics

Project start date: 1 April 2009
Project duration: 36 months

Deliverable D.1.1 - System Specifications and the Corpus Needs (v2)

Dissemination Level: **Public**

Contents

Title	System Specifications and the Corpus Needs
Type	Minor Deliverable
Ref	D.1.1
Target Version	V 2
Current Version	V 2
Author(s)	Gregorio Martínez Ruíz (CRIC)
Reviewer(s)	All partners
Approver(s)	Al partners
Approval date	End September 2010
Release date	End September 2010

1. Introduction

This deliverable sets up the technical specifications (or goals) of the SignSpeak system, along with the context-domain where the translations are going to take place.

2. System specifications

The specifications of the system are listed underneath:

1. **Multimodal system.** Signed languages involve many simultaneous channels for communicating, mainly both hands, face expressions and head movements. SignSpeak seeks to explicitly exploit the complementarities and redundancies between these communication channels, especially in terms of boundary detection. For signed language recognition and translation, SignSpeak will consider the dominant and non-dominant hand, along with head nodding as a non-manual feature for identifying negation. Quantitative measurements of other non-manual gestures (eyes and mouth) will be possible with regard to WER (word error rate) in signed recognition, but detailed and time-consuming annotations of eye and mouth aperture are of limited interest.
2. **More natural.** The signer will speak without wearing gloves or other types of sensors or markers. The entire process will be vision based (non-invasive system) using standard (web) cameras allowing for natural signing with greater acceptance by the deaf community.
3. **Robustness and self-adaptation to the changing ambient conditions**. During the project, research will target the development of detection and tracking techniques to increase robustness with respect to ambient conditions, including signer and clothing independence (see Point 4), viewpoint and lighting variations, and transient occlusions and minor background clutter, as illustrated in the pictures below. Where needed, specific additional recordings will be made for testing the functioning of the system under different ambient conditions.

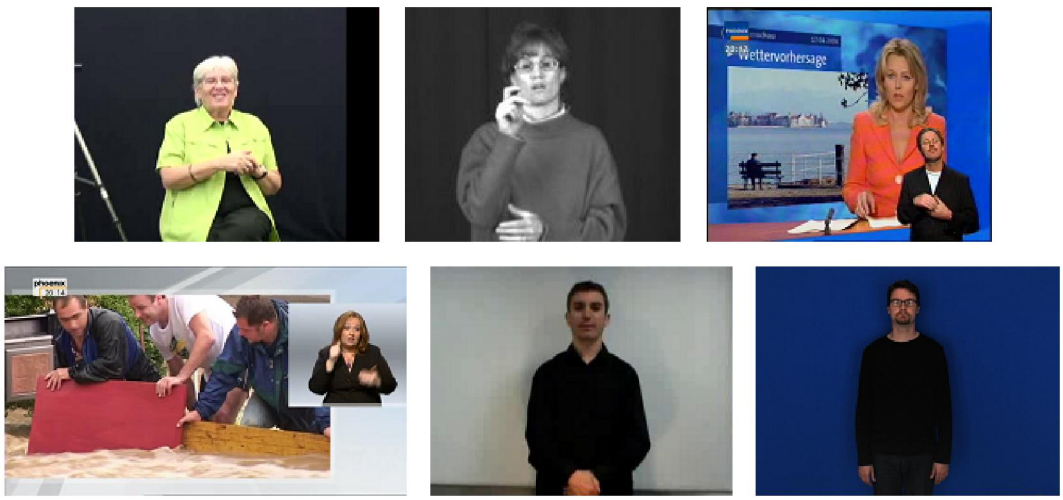


Figure 1. Ambient conditions of the corpora used for developing SignSpeak.

4. **Signer-dependency vs. Signer-independency**. The main goal within the SignSpeak project is to develop a signer-dependent system; signer independency is a long-term goal (beyond the end of the project), which can also be reached thanks to

the statistical approach and the usage of speaker adaptation techniques for gesture and sign language recognition. For speech recognition, the system will be more reliable for more words by training SignSpeak with a concrete signer than for working with a random signer.

5. **Contextual translation.** The system will carry out continuous sign language translation within a context, not merely identifying isolated signs.
6. **Multilingual.** One scientifically challenging task is that there are many different sign languages in Europe, with only a few described grammars. The suggested recognition and translation systems will be based on statistical methods for modelling the appearance and the grammar: these methods have proven to be the most powerful techniques for automatic speech recognition and machine translation in the last years. In addition, the advantages of using these data driven methods gives the technology robustness and scalability to other languages by using different training data. Therefore, although the project will be developed to work with NGT, the system will be also trained and tested to smaller extent in German Sign Language (DGS) and maybe in American Signed Language and Irish Sign Language (it depends on the size of the Corpora available).
7. **Spatial Reference Handling is not considered in the SignSpeak project.** This refers to the analysis of the spatial information containing the entities created during the sign language discourse. While difficult to extract, its analysis would bear new possibilities for the translation, since it could reduce the ambiguity of words that are typically a problem in translation systems (e.g. pronouns). This is too challenging an objective for a three-year project and therefore is not considered as a SignSpeak objective.
8. **Software Integration.** The different prototypes developed separately for multimodal visual analysis, sign language recognition and translation will be integrated by communicating the different applications under a common framework, as shown in next figure.

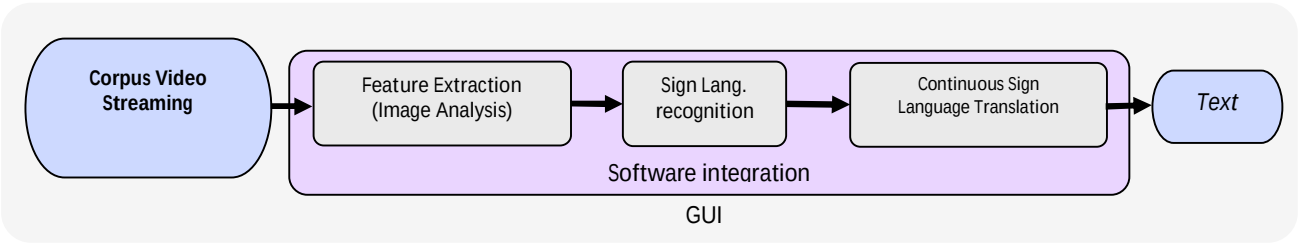


Figure 2. SignSpeak framework

A graphical user interface (GUI) will be designed and developed to monitor inputs-outputs of each subsystem, and to control the parameters involved in the functioning of the system. The GUI will be carried out on WP6.

9. **Context-domain of the translations**. For the Sign Language of the Netherlands, SignSpeak works with video records (Corpus-NGT) created by posing 15 questions to 46 pairs of signers; these questions elicit ‘discussions’ about issues related to the deaf community and deafness. After analysing the observations (word-frequency) in the Corpus NGT (deliverable D.1.2 “Nature of available NGT corpora (ECHO and CNGT)”), it has been selected this 'discussion' domain for targeting the SignSpeak translations.

On the other hand, for demonstrating that SignSpeak is a multilingual system, to a smaller extent we are going to train and test the system in German Sign Language (DGS); in this case, a smaller corpus is built up by recording the weather forecast in a German TV-station; therefore, the context domain is going to be the weather forecast.

10. **Real time factor around 20 for translating NGT**. It is not going to be a real time demonstrator. A real time factor of 20 means that 6 seconds of video records will take 2 minutes for providing the translation. An online demonstration is foreseen for translating the sign language of The Netherlands (NGT), in contrast to the other focused sign language (DGS), where the demonstration will be done by offline evaluations due to the smaller size of the Corpora available for training the system.
11. **Vocabulary size** around 4.000 words for NGT; younger signers (bellow 50 years) will be targeted for reducing generational variations, and from Northern region (largest part of the Corpus NGT) for reducing regional variations. That means a total 10 hours of annotated video records.

3. Corpus Needs

For obtaining a good performance of SignSpeak system, a video corpora for training the system is necessary, in which the total number of words (also called 'token') are repeated several times for each lexicon word (called 'types'). The current state of the Corpus NGT has been annotated by a group of about 10 deaf research assistants, and the annotation guidelines have been under constant development during this period (2006-2010). Currently all existing files are being revised to match the current standards. When this process is done, annotator consistency will be evaluated in the second year of the project for the two annotators that work for SignSpeak

For sign language recognition the data needed is:

- Gloss annotations at sentence level with sentence boundaries; no gloss boundaries.
- Running words per vocabulary entry on average (Token/Types ratio): 15 observations
- Singletons (words with only one observation) < 40%
- Words per sentence: 5-15 words.

Whereas, for signed language translation the data needed is:

- Bilingual annotations: NGT/Dutch, DGS/German, ...
- Running words per vocabulary entry on average (Token/Types ratio): 20 observations
- Singletons (words with only one observation) < 40%
- Words per sentence: 5-12 words.

The technical justification given for these values are presented in [1] and [2].

[1] P. Dreuw, H. Ney, G. Martinez, O. Crasborn, J. Piater, J. Miguel Moya, and M. Wheatley. The SignSpeak Project - Bridging the Gap Between Signers and Speakers. In International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, May 2010.

[2] J. Forster, D. Stein, E. Ormel, O. Crasborn, and H. Ney. Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation. In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010.

As said previously, SignSpeak is being developed using different corpora: Boston, Phoenix and NGT are in American, German and Dutch signed languages, respectively. The following table summarises their features at the beginning and at the end of the project:

	BOSTON-104	Phoenix		Corpus-NGT	
year	2007	2009	2011	2009	2011
recordings	201	78	400	116	300
running words	0.8k	10k	50k	30k	80k
vocabulary size	0.1k	0.6k	< **2.5k**	3k	< **5k**
T/T ratio	8	15	> **20**	10	> **20**

Figure 3. Expected corpus annotation progress of the RWTH-PHOENIX and Corpus-NGT corpora.

Different subsets of the same corpora will be defined for evaluating SignSpeak in different areas:

Corpus	Evaluation Areas			
	Isolated Recog.	Continuous Recog.	Tracking	Translation
Corpus-NGT	✓	✓	✓	✓
RWTH-BOSTON-50	✓	✖	✓	✖
RWTH-BOSTON-104	✖	✓	✓	✖
RWTH-BOSTON-400	✖	✓	✖	✖
RWTH-PHOENIX-v1.0	✓	✓	✷	✓
RWTH-PHOENIX-v2.0	✖	✓	✷	✓
ATIS-ISL	✖	✓	✓	✓
SIGNUM	✓	✓	✷	✖
OXFORD	✖	✖	✓	✖

Figure 4. Freely available sign language corpora and their evaluation areas (✖: unsuitable or unannotated, ✓: already annotated, ✷: annotations underway)

4. Specifications of the communication architecture

Final possible applications of the SignSpeak technology have been identified when SignSpeak proposal was submmitted. These applications have been considered during the set up of the specifications presented previously; a more complete use case analysis of Signspeak applications wil be carried out in WP9.

In task 1.1, as part of the system specifications, TID has studied the communication protocol between the automatic sign recognition service and the platform that contains the VoIP services. To get this definition the following steps were necessary:

- Study the state of the art regarding PBXs. For this study the most appropriate tool to develop the necessary services will be chosen.
- Identify and build the necessary architecture to communicate the VoIP services with the automatic recognition platform.
- Configure a basic PBX to test communication with our partners.
- Perform bidirectional communication tests to check the viability of the solution.
- Perform communication tests to check the quality of the video calls.
- Improve the quality of the video calls.

- Study and test different webcam models to find the more appropriate for the requirements of the project.

The progress and the results of these tasks were:

- PBX - Study the state of the art

After evaluating several tools, both proprietary and free, it was concluded that the application that best meet the needs of the project was Asterisk, which is a known open source PBX with a great number of options and configuration possibilities, from a consolidated and reliable community of developers supporting the project.

- Identify and build the architecture

To communicate the automatic recognition platform with the services developed above the PBX is necessary to study possible communication problems.

This problem was mainly of visibility between machines and security. Regarding visibility it was quite complicated to avoid NAT, while in the case of security, the definition of the service ports was the hardest task.

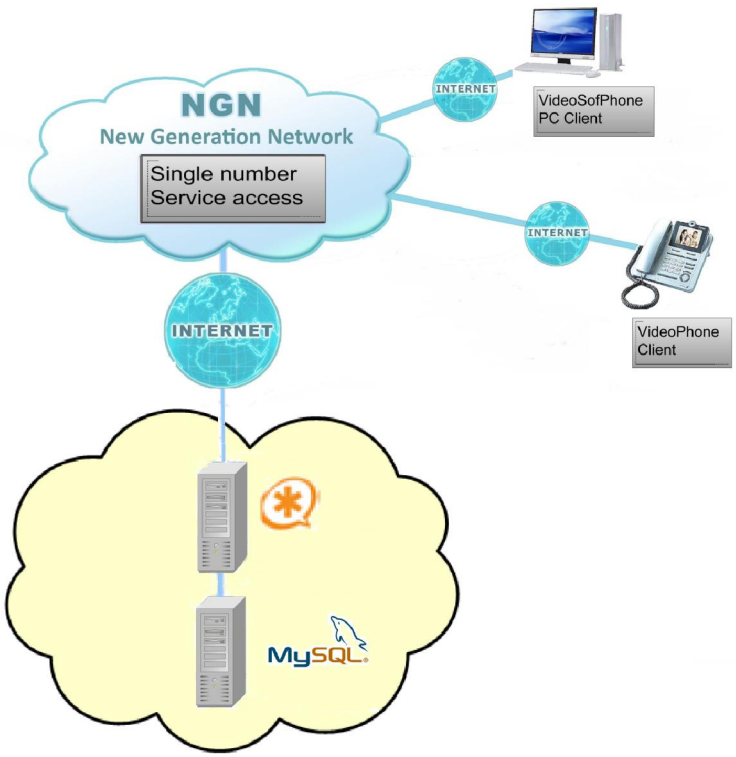


Figure 5. Architecture of the system

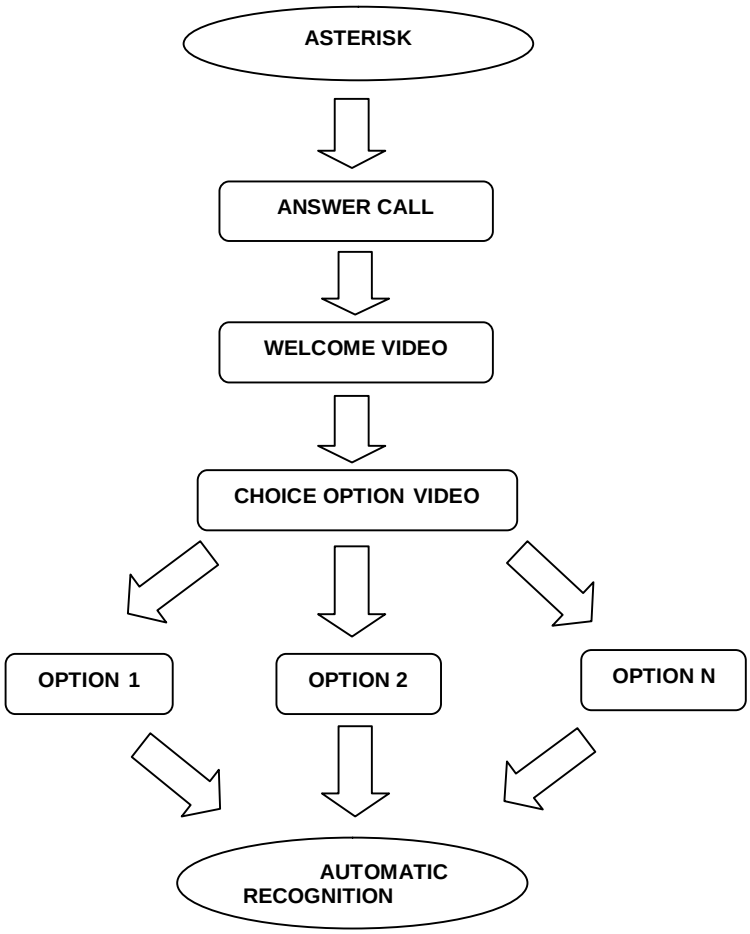
- Configuration of a Basic PBX

Once the architecture was defined, the next step was build a basic PBX to provide the necessary services that allow the end user to call the system, receive information, and access to the automatic sign language tool.

The flow of execution defined for this PBX is the following:

- Answer the call.
- Show the user a video explaining the system in sign language with subtitles.

- Show the user a video asking what option wants to choose (in case there is more than one option).
- Wait for key pulsation, process the key and move the execution of the chosen service.
- Communicate with the recognition system.



- Viability testing

One of the fears after building the architecture was to find problems of communication in any part of the solution. The hardest problem was to find the way to communicate from Asterisk to an account in other country.

To solve this setback, it was necessary to use the NGIN (next generation intelligent network) of Telefónica. This network makes possible the communication, not only between digital and analogical devices, but also between 3G mobiles and VoIP phones. For this reason the architecture of the network and its possibilities in the project were studied.

Finally, the solution was built thanks to the configuration of some accounts of NGN, one on the side of the Asterisk server (in Spain), and other on the side of the phone used for testing (in Holland).

In this context, both video and audio gave excellent results in both directions.

Note: These tests were realized with Radboud University of Nijmegen (RU).

- Quality testing

With the operative platform functioning, one of the worries of the project was the quality of the recorded image. And this set up new inconvenience, the recording of the videos,

because the PBX used has a special format of codification which does not work for standard players. Even worse, the only existing tool to convert videos to/from Asterisk uses the h263 codec, which does not have enough quality for the needs of the project. To get the best qualify in the codification, the project uses the h264 Advanced Video Coding a block-oriented motion-compensation-based codec standard develop by the ITU-T Video Coding Experts Group (VCEG) together with the ISO/IEC Moving Picture Experts Group (MPEG).

After studing different alternatives, finally a tool to convert the received format of the record in a standard format understanding for any player was developed. This work was quite complicated and required great effort, first to study and understand the codec and formats, and later to develop the solution.

The final result were two plug-ins of gstreamer, one to decode the video recorded by Asterisk (with a special payload and rtp packets inside) into avi format video that can be played in standard players; and another to encode from avi format to Asterisk format to use videos recorded by any standard record as an Asterisk video.

Note: These tests were realized with Radboud University of Nijmegen (RU).

- Improve the quality

With the plug-in working correctly, the next step was changing some parameters of the codification to improve the quality of the image, so the bandwidth and frame rate was adjusted, the background of the video was changed to improve the contrast, the position and quantity of light was modified to avoid shadows.

Note: These tests were realized with Radboud University of Nijmegen (RU).

- Study and test different webcam models

To get the best qualify in the reception of the image was the main reason to start an intense study of the properties of the webcams available for the project.

After comparing and contrasting several devices of a huge range of prices and brands (Philips, Logitech, Trust, Microsoft), the main conclusion was the importance of the frame rate parameter, above all the possibility accessing this parameter through the cam´s driver. Another relevant parameter is the bandwith, but in most cases this depends more on the quality of the connection to the network. So, a good choice would be a webcam with at least 30 frames per second, and 256 kbs of bandwidth.

The last part of the test was dedicated High Definition. For this test a Microsoft HD webcam capable of offering a quality of 720p was purchased. For this test it was necessary to modify the codec used for codification/decodification.

The result of the HD test shows high hardware requirements, not only regarding to the capture devices, but also for the machine where the softphone is installed, and even for the network´s quality. To sum up, a speedy processor, a good RAM memory, and a connection with an excellent bandwidth are necessary to reap the benefits of the HD devices.

Furthermore, there are other requirements, such as the need for a complete corpus that includes a wide range of recorded samples with this kind of webcam, and adapting the system to the use of this device.

5. Conclusions

The specifications of the SignSpeak system have been finalised and the corpus needs have been identified.