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.

Dissemination Level: **Public**

Contents

1. Introduction

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

2. System specificatios

The specifications of the system are listed underneath:

1. **Multimodal system**. Due to the many simultaneous ‘channels’ of signed languages (two hands, face, head, upper body), the system will extract information not only from the dominant hand, but also from the non-dominant hand and from the facial expression and body position (shoulders, elbows and chest). SignSpeak seeks to explicitly exploit the complementarities and redundancies between these communication channels, especially in terms of boundary detection. This will allow a self-assessment of its own performance, which should yield a high level of robustness to subsystem failures and graceful behaviour in unforeseen circumstances. More details are expounded in WP3 description.
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 be targeted at the development developing robust feature extraction techniques: the hands are signing often in front of the face (occlusions), and standard face detection methods often fail due to strong facial expressions, head tilt and head turns: that is a challenging task in sign language recognition. Additional research will be carried out to allow the system to work independently of the background colour and the signers’ clothes and brightness, in order to enable robust tracking and speed measurements of the targeted body parts.
4. **Signer-independency**. Thanks to the statistical approach for gesture and sign language recognition, the system will be gender and age-independent similar to robust automatic speech recognition systems. Signer independence also implies **pronunciation, language modelling adaptation and the usage of speaker adaptation techniques**. Due to dialects in natural continuous sign language, signs with the same meaning often differ significantly in their visual appearance and in their duration. In addition, the clouthing are not going to be controlled (just avoid white clouthing).
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 Irish Sign Language (it depends on the size of the Corpora available).

7. **Spatial Reference Handling.** A challenging task will be to analyse the spatial information containing the entities created during the sign language discourse. While difficult to extract, its analysis, it also bears new possibilities for the translation, since it could reduce the ambiguity of words that are typically a problem in translation systems (e.g. pronouns). References in signing space occur quite often to refer to previously deposited objects in the virtual signing space.
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. A graphical user interface will be designed and developed for the easy use of the system.
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 a test the system in German Sign Language (DGS); in this case, a smaller corpora 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 translationing 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** (for NGT) around 4.000 words.

3. Corpus Needs

For obtaining a good performance of SignSpeak system, it is necessary a video corpora for training the system where the total number of words (also called ‘token’) are repeated several times (what is called ‘types’).

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.

As said previously, in SignSpeak we are working with two corpora, one for the Sign Language of the Netherlands (Corpus-NGT), and a smaller one for German Sign Language (RWTH-Phoenix). In order to increase the number of observations, additional videos will be recorded an annotated.

In the case of Corpus-NGT, with the aim of obtaining the targeted words, two solutions have been identified for recording the new data in the 'discussion' domain explained in the previous point:

- By short questions triggering NGT signers to give an opinion related to the deaf community and deafness.
- By showing short fragments of discussions to multiple NGT signers and to ask them to repeat it.

In the case of corpus RWTH-Phoenix, additional videos will be recorded from the German TV-station.

The current and the foreseen features of the corpora are gathered in next table:

	RWTH-Phoenix		Corpus-NGT	
	2009	2011	2009	2011
recordings	78	400	116	300
running words	10k	50k	30k	80k
vocabulary size	0.6k	2.5k	3k	> 5k ?*
T/T ratio	15	20	10	< 20 ?*

* Foreseen values to be confirmed.

The new records will be carried out under task 1.5 "Create an additional NGT corpora with the extra features needed".

4. Conclusions

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