

# Scientific understanding and vision-based technological development for continuous sign language recognition and translation – [www.signspeak.eu](http://www.signspeak.eu) – FP7-ICT-2007-3-231424

## - Annual Public Report -

### **Abstract**

The SignSpeak project will be the first step to approach sign language recognition and translation at levels already obtained in similar technologies such as automatic speech recognition or statistical machine translation of spoken languages. Deaf communities revolve around sign languages as they are their natural means of communication. Although deaf, hard of hearing and hearing signers can communicate without problems amongst themselves, there is a serious challenge for the deaf community in trying to integrate into educational, social and work environments. The overall goal of SignSpeak is to develop a new vision-based technology for recognizing and translating continuous sign language to text. New knowledge about the nature of sign language structure from the perspective of machine recognition of continuous sign language will allow a subsequent breakthrough in the development of a new vision-based technology for continuous sign language recognition and translation. Existing and new publicly available corpora will be used to evaluate the research progress throughout the whole project.

### **Introduction**

The SignSpeak project is one of the first EU funded projects that tackles the problem of automatic recognition and translation of continuous sign language; it is a 3 year project, starting on 1<sup>st</sup> of April 2009. The overall goal of the SignSpeak project is to develop a new vision-based technology for recognizing and translating continuous sign language (i.e. provide Video-to-Text technologies), in order to provide new e-Services to the deaf community and to improve their communication with the hearing people. The current rapid development of sign language research is partly due to advances in technology, including of course the spread of Internet, but especially the advance of computer technology enabling the use of digital video. The main research goals are related to a better scientific understanding and vision-based technological development for continuous sign language recognition and translation:

- understanding sign language requires better linguistic knowledge.
- large vocabulary recognition requires more robust feature extraction methods and a modeling of the signs at a sub-word unit level.
- statistical machine translation requires large bilingual annotated corpora and a better linguistic knowledge for phrase-based modeling and alignment.

Therefore, the SignSpeak project combines innovative scientific theory and vision-based technology development by gathering novel linguistic research and the most advanced techniques in image analysis, automatic speech recognition (ASR) and statistical machine translation (SMT) within a common framework.

### **SignSpeak Specifications**

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 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 of 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 clothing is not going to be controlled (just avoid white clothing).
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 a 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, accounting around 90 hours of clips; these questions elicit 'discussions' about issues related to the deaf community and deafness. After analysing the observations (word-frequency) in the Corpus NGT, 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** (for NGT) around 4.000 words.

## **Research and Challenges in Automatic Sign Language Recognition**

In the following points it is briefly discussed the most important topics to build up a large vocabulary sign language recognition system.

### **Languages and Available Resources.**

Almost all publicly available resources, which have been recorded under lab conditions for linguistic research purposes, have in common that the vocabulary size, the types/token ratio (TTR), and signer/speaker dependency are closely related to the recording and annotation costs. Data-driven approaches with systems being automatically trained on these corpora do not generalize very well, as the structure of the signed sentences has often been designed in advance, or offer small variations only, resulting in probably overfitted language models. Additionally, most self-recorded corpora consist only of a limited number of signers.

In the recently very active research area of sign language recognition, a new trend towards broadcast news or weather forecast news can be observed. Due to limited preparation time of the interpreters, the grammatical differences between “real-life” sign language and the sign language used in TV broadcast (being more close to Signed Exact English (SEE)) are often significant.

### **Environment Conditions and Feature Extraction.**

Further difficulties for such sign language recognition frame works arise due to different environment assumptions. Most of the methods developed assume closed-world scenarios, e.g. simple backgrounds, special hardware like data gloves, limited sets of actions, and a limited number of signers, resulting in different problems in sign language feature extraction or modeling.

### **Modeling of the Signs.**

In continuous sign language recognition, as well as in speech recognition, co-articulation effects have to be considered. One of the challenges in the recognition of continuous sign language on large corpora is the definition and modelling of the basic building blocks of sign language. The use of whole-word models for the recognition of sign language with a large vocabulary is unsuitable, as there is usually not enough training material available to robustly train the parameters of the individual word models. A suitable definition of sub-word units for sign language recognition would probably alleviate the burden of insufficient data for model creation.

In ASR, words are modelled as concatenated sub-word units. These sub-word units are shared among the different word-models and thus the available training material is distributed over all word-models. On the one hand, this leads to better statistical models for the sub-word units, and on the other hand it allows recognizing words which have never been seen in the training procedure using lexica. According to previous studies, a phonological model for sign language can be defined, dividing signs into their four constituent visemes, such as the hand shapes, hand orientations, types of hand movements, and body locations at which signs are executed. Additionally, non-manual components like facial expression and body posture are used. However, no suitable decomposition of words into sub-word units is currently known for the purposes of a large vocabulary sign language recognition system (e.g. a grapheme-to-phoneme like conversion and use of a pronunciation lexicon). The most important of these problems are related to the lack of generalization and overfitting systems, poor scaling and unsuitable databases for mostly data driven approaches.

## Speech and Sign Language Recognition

Automatic speech recognition (ASR) is the conversion of an acoustic signal (sound) into a sequence of written words (text).

Due to the high variability of the speech signal, speech recognition – outside lab conditions – is known to be a hard problem. Most decisions in speech recognition are interdependent, as word and phoneme boundaries are not visible in the acoustic signal, and the speaking rate varies. Therefore, decisions cannot be drawn independently but have to be made within a certain context, leading to systems that recognize whole sentences rather than single words.

One of the key ideas in speech recognition is to put all ambiguities into probability distributions (so called stochastic knowledge sources, see Figure 1).

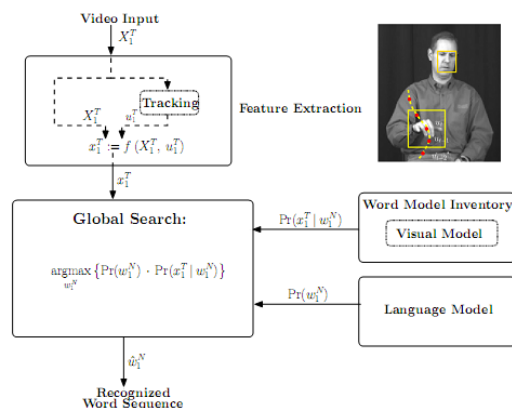


Figure 1. Sign language recognition system overview

Then, by a stochastic modelling of the phoneme and word models, a pronunciation lexicon and a language model, the free parameters of the speech recognition framework are optimized using a large training data set. Finally, all the interdependencies and ambiguities are considered jointly in a search process which tries to find the best textual representation of the captured audio signal. In contrast, rule-based approaches try to solve the problems more or less independently.

In order to design a speech recognition system, four crucial problems have to be solved:

1. pre-processing and feature extraction of the input signal,
2. specification of models and structures for the words to be recognized,
3. learning of the free model parameters from the training data, and
4. search the maximum probability over all models during recognition (see Figure 1).

**Differences Between Spoken Language and Sign Language.** Main differences between spoken language and sign language are due to linguistic characteristics like simultaneous facial and hand expressions, references in the virtual signing space and grammatical differences:

- Simultaneousness: a signer can use different communication channels (facial expression, hand movement, and body posture) in parallel.
- Signing Space: entities like persons or objects can be stored in a 3D body-centered space around the signer, by executing them at a certain location and later just referencing them by pointing to the space – the challenge is to define a model for spatial information handling.
- Coarticulation and Epenthesis: In continuous sign language recognition, as well as in speech recognition, coarticulation effects have to be considered. Due to location changes in the 3D signing space, we also have to deal with the movement epenthesis problem. Movement epenthesis refers to movements which occur regularly in natural sign language in order to move from the end state of one sign to the beginning of the next one. Movement epenthesis conveys no meaning in itself but contributes phonetic information to the perceiver.
- Silence: opposed to automatic speech recognition, where usually the energy of the audio signal is used for the silence detection in the sentences, new spatial features and models will have to be defined for silence detection in sign language recognition. Silence cannot be detected by simply

analyzing motion in the video, because words can be signed by just holding a particular posture in the signing space over time. Further, the rest position of the hand(s) may be somewhere in the signing space.

## Sign Language Translation

The goal of machine translation (MT) is the translation of a text given in some natural source language into a natural target language. The input can be either a written sentence or a spoken sentence that was recognised by a speech recognition system. Statistical methods, similar to those used in speech recognition, describe the structure of the sentences of the target language, the language model, and the dependencies between words of the source and the target language, the translation model.

Sign languages have a unique grammar and vocabulary that are independent of spoken languages. SignSpeak will implement a statistical sign language machine translation system (SMT). Existing methods for sign languages suffer from two main limitations. Rule-based approaches are inflexible in their domain because they require heavy linguistic rules and definitions, which cannot be adapted to other domains or other languages without great cost. Corpus-based approaches suffer from data sparseness, so that results only give preliminary directions and the statistic significance is often doubtful. SignSpeak will progress beyond the state of the art by working in complex and continuous sign language scenarios. Our system will be context-dependent, taking into account preceding and following signs and their location within the signing space. Another challenge is to model the reordering (see Figure 2). Since SMT does not rely on rules that need to be defined externally, it can be easily tuned to new domains and languages assuming a reasonably-sized data set are available, resolving both the problems of data sparseness and lack of flexibility.

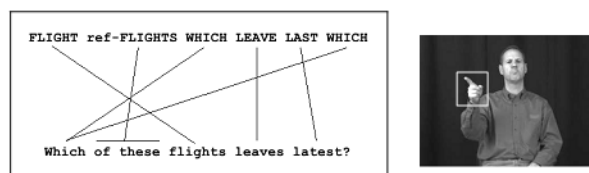


Figure 2: Different word orderings and pointing events have to be handled in sign language translation

## Towards a Speech-to-Speech Translation System

The interpersonal communication problem between signer and hearing community could be resolved by building up a new communication bridge integrating components for sign-, speech-, and text-processing. To build a sign-to-speech translator for a new language, a six component-engine must be integrated (see Figure 3), where each component is in principle language independent, but requires language dependent parameters/models. The models are usually automatically trained but require large annotated corpora. In SignSpeak, a theoretical study will be carried out about how the new communication bridge between deaf and hearing people could be built up by analyzing and adapting the ASLR and MT components technologies for sign language processing.

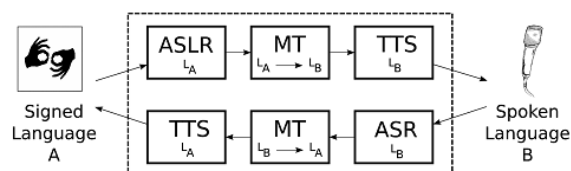


Figure 3. Complete six components-engine necessary to build a Sign-To-Speech system (components: automatic sign language recognition (ASLR), automatic speech recognition (ASR), machine translation (MT), and text-to-speech/sign (TTS)).



Once the different modules are integrated within a common communication platform, the communication could be handled over 3G phones, media center TVs, or video telephone devices. The following application scenarios would be possible:

- e-learning of sign language
- automatic transcription of video e-mails, video documents, or video-SMS
- video subtitling and annotation

The novel features of such systems provide new ways for solving industrial problems. The technological breakthrough of SignSpeak will clearly impact on other applications fields:

- Improving human-machine communication by gesture: vision-based systems are opening new paths and applications for human-machine communication by gesture, e.g. Play Station's EyeToy or Microsoft Xbox's Natal Project, which could be interesting for physically disabled individuals or even blind people as well.
- Medical sector: new communication methods by gesture are being investigated to improve the communication between the medical staff, the computer, and other electronic equipments. Another application in this sector is related to web- or video-based e-Care / e-Health treatments, or an auto-rehabilitation system which makes the guidance process to a patient during the rehabilitation exercises easier.
- Surveillance sector: person detection and recognition of body parts or dangerous objects, and their tracking within video sequences or in the context of quality control and inspection in manufacturing sectors.