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

– www.signspeak.eu – FP7-ICT-2007-3-231424

– Annual Public Report – 2nd year –

Abstract

The SignSpeak project is the first step in taking sign language recognition and translation to levels already obtained in automatic speech recognition or statistical machine translation of spoken languages. Deaf communities revolve around sign languages as their natural means of communication. Although 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 (i.e. provide Video-to-Text technologies), in order to provide new e-Services to the deaf community and improve their communication with hearing people. New knowledge about the nature of sign language structure from the perspective of machine recognition of continuous sign language will lead to a 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 to tackle the problem of automatic recognition and translation of continuous sign language; it is a 3 year project, started on the 1st of April 2009. 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 an improvement of linguistic knowledge.
- large vocabulary recognition requires more robust feature extraction methods and modeling of the signs at a sub-word unit level.
- statistical machine translation requires large bilingual annotated corpora and an advanced linguistic knowledge for phrase-based modelling 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**. Signed languages involve many simultaneous channels for communicating, mainly both hands, face expressions and head movements. SignSpeak seeks to exploit the complementarities and redundancies between these communication channels, especially in terms of boundary detection. For signed language recognition and translation, SignSpeak considers the

dominant and non-dominant hand, along with head shaking 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, regardless of the signer and their clothing (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.

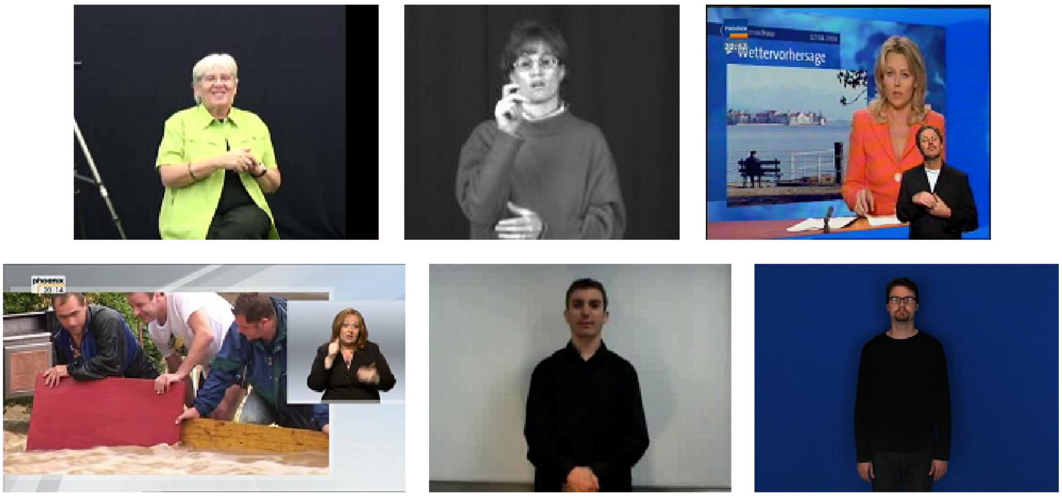


Figure 1: Ambient conditions of the corpora used for developing SignSpeak

4. **Signer-dependency vs. Signer-independency**. The main goal of 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, but 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, using these data driven methods gives the technology robustness and scalability to other languages by using different training data. Thus, although the project will be developed to work with NGT, the system will be also trained and tested to a lesser extent in German Sign Language (DGS) and maybe in American Signed Language and Irish Sign Language, depending on the size of the available Corpora.
7. **Spatial Reference Handling.** 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.

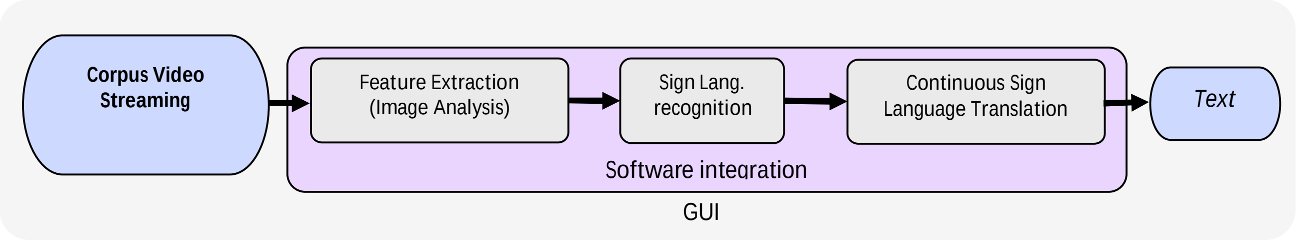

Corpus Video Streaming
Feature Extraction (Image Analysis)
Sign Lang. recognition
Continuous Sign Language Translation
Text
Software integration
GUI

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)"), this 'discussion' domain has been selected for targeting the SignSpeak translations.

   On the other hand, to demonstrate that SignSpeak is a multilingual system, 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, in a more controlled context domain scenario (smaller vocabulary size).

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 to provide 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 (below 50 years of age) 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.

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 over fitted 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 modelling.

Modelling 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 over fitting 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).

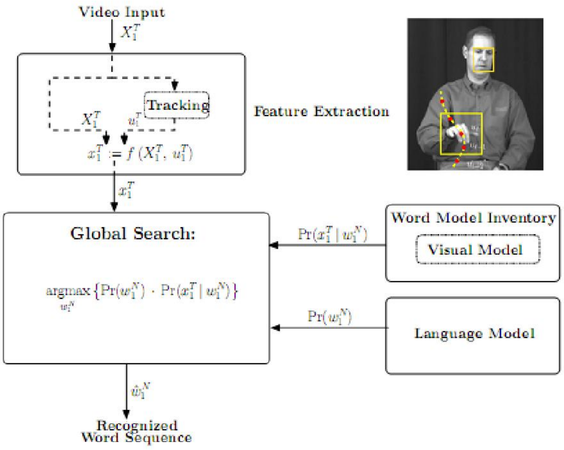


Figure 3. Sign language recognition system overview

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

The 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.

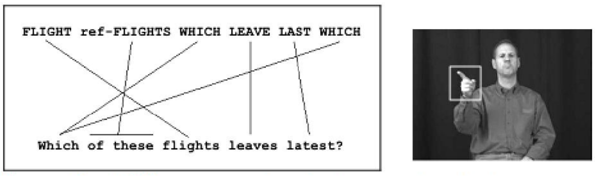

FLIGHT ref-FLIGHTS WHICH LEAVE LAST WHICH
Which of these flights leaves latest?

Figure 4: 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.

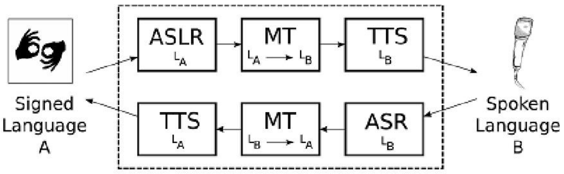


Figure 5. Complete six components-engine necessary to build a Sign-To-Speech system (components: automatic sign language recognition (ASLR), automatic speech recognition (ASR), machine translation (MT), and text-to-speech/sign (TTS)).

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.

Progress and achievements

A conceptual scheme of the work planned is presented in next figure.

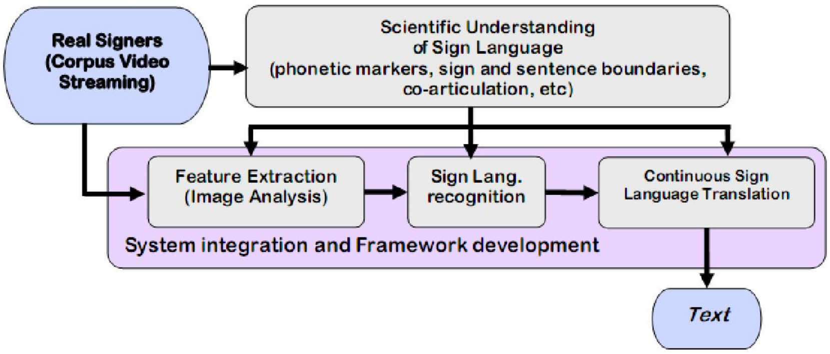

Real Signers (Corpus Video Streaming)
Scientific Understanding of Sign Language (phonetic markers, sign and sentence boundaries, co-articulation, etc)
Feature Extraction (Image Analysis)
Sign Lang. recognition
Continuous Sign Language Translation
System integration and Framework development
Text

Figure 6. Conceptual scheme of the work planned in SignSpeak project

SignSpeak has reached month 21 (December 2010). The following points summarise the progress made in the different blocks:

- **Video Corpus:** The features of the existing NGT Corpora (sign language of The Netherlands, http://www.ru.nl/corpusngt/) have been studied for defining the requirements of gloss, translations, sentence boundaries and non-manual annotations, as well as for selecting additional videos to be recorded to increase the word repetition. Considerable effort has been invested in carrying out all these annotations, which is scheduled until month 24.

  At month 21 the consortium will start enlarging a corpora with videos in German Sign Language (DGS), the RWTH-PHOENIX database, aiming at showing how SignSpeak works by handling databases with different features (different languages, context domain, vocabulary size, recording conditions...).

- **Scientific understanding of signed languages**: Literature study reveals that there are no 'hard' cues thus far for sentence boundaries to be exploited for sign recognition research, and a new approach is required whereby combinations of cues are analyzed. Several of our linguistic studies have shown that some lexical items are predictive for sentence boundaries, given that these items often occur at the start or the end of sentences, although not consistently. Thus far, lexical items and prosodic cues were analyzed separately in past studies. Both have shown to be predictive, however, not sufficiently to be able to detect sentence boundaries. Prosodic cues and lexical cues should therefore be combined to predict sentence boundaries for automatic sign language translations. The literature study revealed that video analyses and linguistic analyses can be mutually informative to gain further insight in the exact phonetic/prosodic cues present at sentence boundaries and should be exploited in further studies.

- **Feature extraction**: a Baseline Prototype for the multimodal visual analysis has been developed integrating hand and face tracking. The hand tracking method intrinsically allows a quantitative characterization of hand shapes; this avenue will be further pursued during the coming months. The face tracking method allows the quantification of certain aspects of facial expressions such as 3D head orientation and eye and mouth apertures. In addition, spatiotemporal features have been extracted by different approaches and will be integrated in next Advanced Prototype of the multimodal visual analysis to be delivered at the end of the second year of the project.

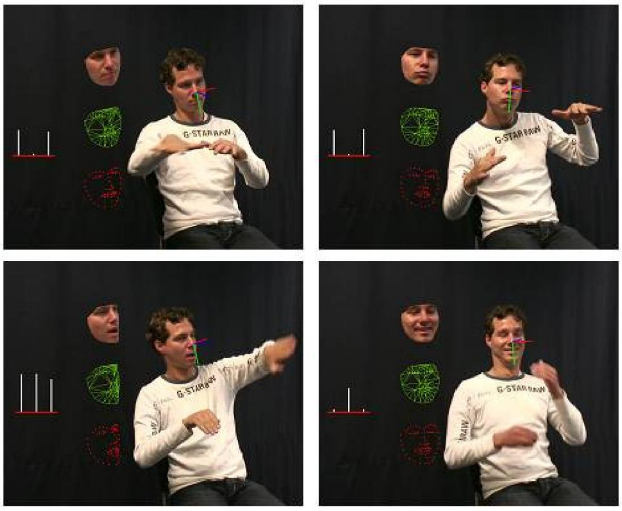


Figure 7. Feature extraction and expression quantification

- **Sign Language Recognition**: a Baseline Prototype for sign language recognition has been developed, which allows the recognition of isolated and continuous sign language data. It offers many configuration possibilities and will allow for the recognition of different signed languages in the future, such as the Corpus-NGT database (Sign Language of the Netherlands or NGT) or the RWTH-PHOENIX-v2.0 database (German Sign Language: GSL). The current prototype has been trained on appearance-based image features only, without any tracking features. The next steps will include the integration of more sophisticated features from the baseline prototype for multimodal visual analysis introduced before.
- **Sign Language Translation to text**: a Baseline Prototype for sign language translation has been developed by employing statistical sign language translation system based on a state-of-the-art hierarchical decoder. It works on continuous sign language, allowing for gaps in the translation by working on a CFG grammar structure. We also tested syntactically motivated methods in pre- and post-processing using additional monolingual data by means of a morpho-syntactic analyser (Morphisto) and a deep syntactic parser (Stanford parser) for German.

Dissemination activities

Partners involved in SignSpeak project have co-organised along with Dicta-Sign partners (another EC funded project working in sign language recognition), two dissemination workshops at the most prestigious conferences in the domain of Language Resources and Computer Vision:

- CSLT 2010: SignSpeak partners RWTH and CRIC were organizing committee member of the “Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010)”, Valletta, Malta, May 22nd-23rd, 2010. Organized as a Language Resources and Evaluation Conference (LREC 2010) post-conference workshop.
  URL: http://www.sign-lang.uni-hamburg.de/lrec2010/cfp.html
- SGA 2010: SignSpeak partners RWTH and ULg were organizing committee member of the “International Workshop on Sign, Gesture, and Activity (SGA 2010)”, Hersonissos, Heraklion, Crete, Greece, Sep 11th, 2010. Organized as a European Conference on Computer Vision (ECCV 2010) satellite workshop.
  URL: http://personal.ee.surrey.ac.uk/Personal/R.Bowden/SGA2010/

A total of 11 papers from the SignSpeak consortium have been accepted at international conferences and workshops; two additional papers have been sent to scientific journals and are waiting for their approval. All the papers have been uploaded on the project website: www.signspeak.eu/en/publications.html

Last but not least, the European Union of the Deaf (EUD) had an active role disseminating the SignSpeak project in the Deaf community. On May 2009, EUD held their annual Workshops in which members come together and learn about developments within the Deaf community, including policy and political changes that have an impact on their membership. It is an occasion in which much information is conveyed and views are exchanged. In short, it is an event where members have a truly open dialogue of experiences and of developments within their organisations and how they are faring in the broader social sector nationally. The EUD Workshop took place in Prague, Czech Republic and was attended at nearly full capacity by member organisations. EUD gave a comprehensive one-hour presentation about SignSpeak and also fielded Q&A session with the members.

Additionally, EUD expounded on its participation in SignSpeak project at the following events:

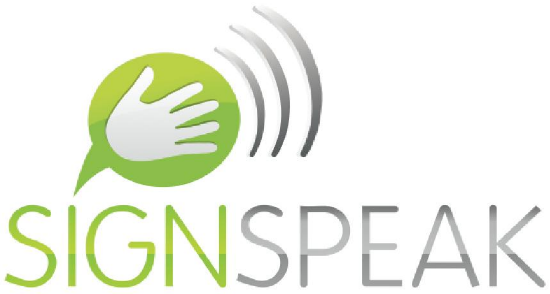
- BSL Charter Conference – Bristol, United Kingdom, April 2009.
- National French Deaf Federation (FNSF) Conference – Limoges, France, May 2009.
- European Federation of Hard of Hearing (EFHOH) Conference – London, United Kingdom, June 2009.
- Bulgarian Deaf Federation 50th Year Anniversary Conference – Sofia, Bulgaria, August 2009.
- European Forum of Sign Language Interpreters Conference (EFSLI) – Tallinn, Estonia, September 2009.
- Liechtenstein Deaf Association Conference – Vaduz, Liechtenstein, November 2009.
- Icelandic Deaf Association 50th Year Anniversary Conference – Reykjavik, Iceland, February 2010.
- World Federation of the Deaf (WFD) Board Meeting – Istanbul, Turkey, March 2010.

SignSpeak logo competition

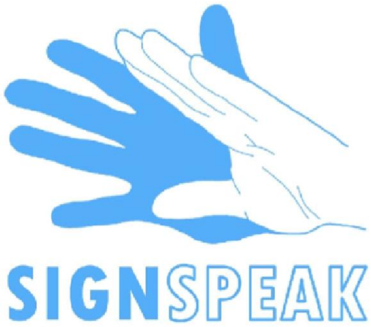
The consortium organised an awarded competition with the aim of both obtaining a new logo for SignSpeak and spreading the goals and scope of the project into the general public but mainly within the Deaf community, as they are the ones who will be benefiting from this new technology. To this end, the advertisement of the competition was published in the SignSpeak and EUD websites, Facebook and other media, encouraging designers to visit the project website to get inspiration.

A total of 17 designers from worldwide have participated in the competition, some of them submitting more than one design. The jury was formed by 12 people from SignSpeak consortium.

The first prize (500€) was for **Carole Langlet** (France):



The 2nd prize (350€) was for **Bart Koolen** (The Netherlands):



The 3rd prize (150€) was for **Alexander Martiyanov** (Russia):