# SIGNSPEAK - BRIDGING THE GAP BETWEEN SIGNERS AND SPEAKERS

*Philippe Dreuw and Hermann Ney*

*Human Language Technology and Pattern Recognition,*
*RWTH Aachen University, D-52056 Aachen*
*Email:* {dreuw,ney}@cs.rwth-aachen.de
*Web:* http://www-i6.informatik.rwth-aachen.de

**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 boundaries 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.

## 1 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.

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.

### 1.1 Sign Languages in Europe

Although sign languages are used by a significant number of people, only a few member states of the European Union (EU) have recognized their national sign language on a *constitutional* level: Finland (1995), Slovak Republic (1995), Portugal (1997), Czech Republic (1998 & 2008), Austria (2005), and Spain (2007). The European Union of the Deaf (EUD)[1] is a European non-profit making organization which aims to establish and maintain EU level dialogue with the "hearing world" in consultation and co-operation with its member National Deaf Associations. The EUD is the only organization representing the interests of Deaf Europeans at European Union level. The EUD has 30 full members (27 EU countries plus Norway, Iceland & Switzerland), and 6 affiliated members (Croatia, Serbia, Bosnia and Herzegovina, Macedonia, Turkey & Israel). Their main goals are the recognition of the right to use an indigenous sign language, the empowerment through communication and information, and the equality in education and employment. In 2008, the EUD estimated about 650,000 Sign Language users in Europe, with about 7,000 official Sign Language Interpreters, resulting in approximately 93 sign language users to 1 sign language interpreter (EUD Survey, 2008). However, the number of sign language users might be much higher, as it is difficult to estimate an exact number – e.g. late-deafened or hard of hearing people who need interpreter services are not always counted as deaf people in these statistics.

### 1.2 Linguistic Research in Sign Languages

Linguistic research on sign languages started in the 1950s, with initial studies of Tervoort [19] and Stokoe [18]. In the USA, the recognition of sign languages as an important linguistic research object only started in the 1970s, with Europe following in the 1980s. Only since 1990, sign language research has become a truly world-wide enterprise, resulting in the foundation of the Sign Language Linguistics Society in 2004[2]. Linguistic research have targeted all areas of linguistics, with the exception of 'phonetics'. The current rapid development of sign language research is partly caused by advances in technology, including of course the spread of the Internet, but especially the advance of computer technology enabling the use of digital video.

Vision-based sign language recognition has only been attempted on the basis of small sets of elicited data (Corpora) recorded under lab conditions (only from one to three signers and under controlled colour and brightness ambient conditions), without the use of spontaneous signing. The same restriction holds for linguistic research on sign languages. Due to the extremely time-consuming work of linguistic annotation, studying sign languages has necessarily been confined to small selections of data. Depending on ones research strategy, researchers either choose to record small sets of spontaneous signing which will then be transcribed to be able to address the linguistic question at hand, or native signer intuitions about what forms a correct utterance.

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

In [15, 22] reviews on research in sign language and gesture recognition are presented. In the following we briefly discuss the most important topics to build up a large vocabulary sign

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

[2] http://www.slls.eu

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 [1], or offer small variations only [24, 7, 4], resulting in probably over-fitted language models. Additionally, most self-recorded corpora consists only of a limited number of signers [21, 2].

In the recently very active research area of sign language recognition, a new trend towards broadcast news or weather forecast news can be observed. The problem of aligning an American Sign Language (ASL) sign with an English text subtitle is considered in [11]. In [3, 6], the goal is to automatically learn a large number of British Sign Language (BSL) signs from TV broadcasts. 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 frameworks 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, coarticulation 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 a 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 to recognize words which have never been seen in the training procedure using lexica. According to the *linguistic* work on sign language by Stokoe [18], 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).

### 1.4 Research and Challenges in Statistical Machine Translation of Sign Languages

While the first papers on sign language translations only date back to roughly a decade [20] and typically employed rule-based systems, several research groups have recently focussed on data-driven approaches. In [17], a SMT system has been developed for German and German sign language in the domain weather reports. Their work describes the addition of pre- and post-processing steps to improve the translation for this language pairing. The authors of [14] have

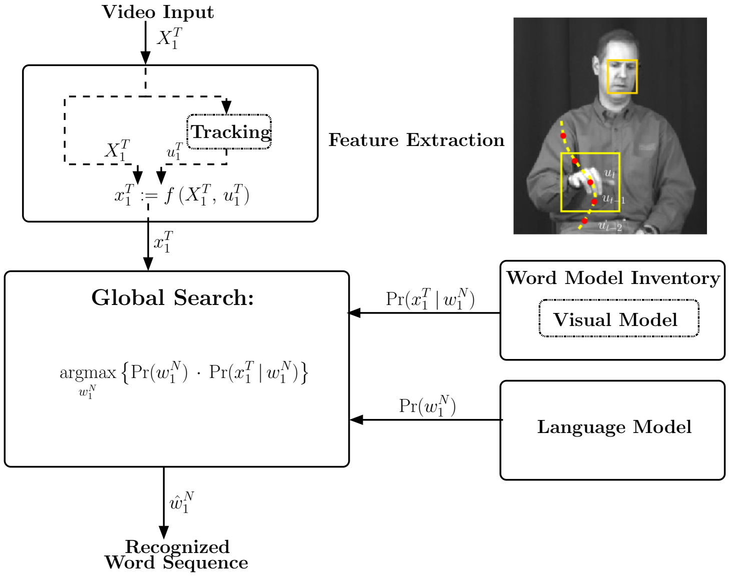

Figure 1: Bayes' decision rule used in ASR and ASLR systems.

explored example-based MT approaches for the language pair English and sign language of the Netherlands with further developments being made in the area of Irish sign language. In [5], a system is presented for the language pair Chinese and Taiwanese sign language. The optimizing methodologies are shown to outperform a simple SMT model. In the work of [16], some basic research is done on Spanish and Spanish sign language with a focus on a speech-to-gesture architecture.

## 2 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 idea in speech recognition is to put all ambiguities into probability distributions (so called stochastic knowledge sources, see Figure 1). 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. preprocessing 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).

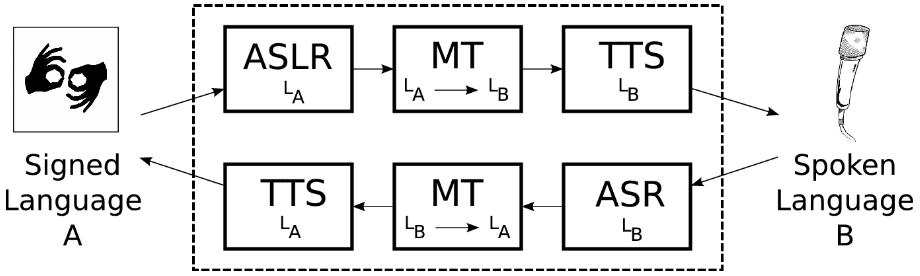

Figure 2: Complete six components-engine necessary to build a Sign-To-Speech system (components: automatic sign language recognition (ASLR), automatic speech recognition (ASR), machine translation (MT), and text-to-speech/sign (TTS))

**Differences Between Spoken Language and Sign Language.** Main differences between spoken language and sign language are due to language characteristics like simultaneous facial and hand expressions, references in the virtual signing space, and grammatical differences as explained more detailed in [9]:

**Simultaneousness:** Major issue in sign language recognition compared to speech recognition – 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 – challenging 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 have to deal with the movement epenthesis problem [21, 23]. Movement epenthesis refers to movements which occur regularly in natural sign language in order to change the location in signing space. Movement epenthesis conveys no meaning in itself but rather changes the meaning of succeeding signs.

**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.

## 3 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 2), 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.

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

The novel features of such systems provide new ways for solving industrial problems. The technological breakthrough of SignSpeak will clearly impact in 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[3], 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 with the computer and other electronic equipments. Another application in this sector is related to web- or video-based *e-Care / e-Health* treatments.

**Surveillance sector:** person detection and recognition of body parts or dangerous objects, and their tracking within video sequences.

## 4 Experimental Results and Requirements

In order to build a Sign-To-Speech system, reasonably sized corpora have to be created for the data-driven approaches. For a limited domain speech recognition task as e.g. presented in [12], systems with a vocabulary size of up to 10k words have to trained with at least 700k words to obtain a reasonable performance, i.e. about 70 observations per vocabulary entry. Similar values must be obtained for a limited domain translation task as e.g. presented in [13].

Similar corpora statistics can be observed for other ASR or MT tasks. The requirements for a sign language corpus suitable for recognition and translation can therefore be summarized as follows:

- annotations should be domain specific (i.e. broadcast news, or weather forecasts, etc.)
- for a vocabulary size smaller than 4k words, each word should be observed at least 20 times
- the singleton ratio should ideally stay below 40%

Existing corpora must be extended to achieve a good performance w.r.t. recognition and translation. During the SignSpeak project, the existing RWTH-Phoenix corpus [17] will be extended to meet these demands (see Table 1).

For automatic sign language recognition, promising results have been achieved for continuous sign language recognition under lab conditions [1, 8]. Even if the performances of the automatic learning approaches presented in [11] and [3, 6] are still quite low, they represent an interesting approach for further research.

[3] http://www.xbox.com/en-US/live/projectnatal/

Table 1: Expected corpus annotation progress of the RWTH-Phoenix corpus in comparison to the limited domain IWSLT corpus.

| | RWTH-Phoenix | | IWSLT |
|---|---|---|---|
| year | 2009 | 2011 | |
| recordings | 78 | 400 | - |
| running words | 10k | 50k | 200k |
| vocabulary size | 0.6k | 2.5k | 10k |
| T/T ratio | 15 | 20 | 20 |

For the task of sign-to-speech recognition and translation, promising results on the publicly available benchmark database RWTH-BOSTON-104 have been achieved for automatic sign language recognition [8] and translation [9, 10] that can be used as baseline reference for other researchers. However, the preliminary results on the larger RWTH-BOSTON-400 database show the limitations of the proposed framework and the need for better visual features, models, and corpora [7].

## References

[1] Agris, U. von and K.-F. Kraiss: *Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition*. In *Gesture in Human-Computer Interaction and Simulation*, Lisbon, Portugal, May 2007.

[2] Bowden, R., D. Windridge, T. Kadir, A. Zisserman and M. Brady: *A Linguistic Feature Vector for the Visual Interpretation of Sign Language*. In *ECCV*, vol. 1, pp. 390–401, 2004.

[3] Buehler, P., M. Everingham and A. Zisserman: *Learning sign language by watching TV (using weakly aligned subtitles)*. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009.

[4] Bungeroth, J., D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way and L. van Zijl: *The ATIS Sign Language Corpus*. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008.

[5] Chiu, Y.-H., C.-H. Wu, H.-Y. Su and C.-J. Cheng: *Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis*. IEEE Trans. PAMI, **29**(1):28–39, 2007.

[6] Cooper, H. and R. Bowden: *Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition*. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009.

[7] Dreuw, P., C. Neidle, V. Athitsos, S. Sclaroff and H. Ney: *Benchmark Databases for Video-Based Automatic Sign Language Recognition*. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008.

[8] Dreuw, P., D. Rybach, T. Deselaers, M. Zahedi and H. Ney: *Speech Recognition Techniques for a Sign Language Recognition System*. In *ICSLP*, Antwerp, Belgium, Aug. 2007.

[9] DREUW, P., D. STEIN, T. DESELAERS, D. RYBACH, M. ZAHEDI, J. BUNGEROTH and H. NEY: *Spoken Language Processing Techniques for Sign Language Recognition and Translation*. Technology and Dissability, 20(2):121–133, June 2008.

[10] DREUW, P., D. STEIN and H. NEY: *Enhancing a Sign Language Translation System with Vision-Based Features*. In *Intl. Workshop on Gesture in HCI and Simulation 2007*, pp. 18–19, Lisbon, Portugal, May 2007.

[11] FARHADI, A. and D. FORSYTH: *Aligning ASL for statistical translation using a discriminative word model*. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, USA, 2006.

[12] KANTHAK, S., A. SIXTUS, S. MOLAU, R. SCHLÜTER and H. NEY: *Fast Search for Large Vocabulary Speech Recognition*, chap. "From Speech Input to Augmented Word Lattices", pp. 63–78. Springer Verlag, Berlin, Heidelberg, New York, July 2000.

[13] MAUSER, A., R. ZENS, E. MATUSOV, S. HASAN and H. NEY: *The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation*. In *International Workshop on Spoken Language Translation*, pp. 103–110, Kyoto, Japan, Nov. 2006. Best Paper Award.

[14] MORRISSEY, S. and A. WAY: *An Example-based Approach to Translating Sign Language*. In *Workshop in Example-Based Machine Translation (MT Summit X)*, pp. 109–116, Phuket, Thailand, 2005.

[15] ONG, S. and S. RANGANATH: *Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning*. IEEE Trans. PAMI, 27(6):873–891, June 2005.

[16] SAN-SEGUNDO, R., R. BARRA, L. F. D'HARO, J. M. MONTERO, R. CÓRDOBA and J. FERREIROS: *A Spanish Speech to Sign Language Translation System for assisting deaf-mute people*. In *ICSLP*, Pittsburgh, PA, 2006.

[17] STEIN, D., J. BUNGEROTH and H. NEY: *Morpho-Syntax Based Statistical Methods for Sign Language Translation*. In *11th EAMT*, pp. 169–177, Oslo, Norway, June 2006.

[18] STOKOE, W., D. CASTERLINE and C. CRONEBERG: *A Dictionary of American Sign Language on Linguistic Principles*. Gallaudet College Press, Washington D.C., USA, 1965.

[19] TERVOORT, B.: *Structurele analyse van visueel taalgebruik binnen een groep dove kinderen*, 1954.

[20] VEALE, T., A. CONWAY and B. COLLINS: *The Challenges of Cross-Modal Translation: English to Sign Language Translation in the ZARDOZ System*. Journal of Machine Translation, 13, No. 1:81–106, 1998.

[21] VOGLER, C. and D. METAXAS: *A Framework for Recognizing the Simultaneous Aspects of American Sign Language*. Computer Vision & Image Understanding, 81(3):358–384, Mar. 2001.

[22] Y. WU, T. H.: *Vision-based gesture recognition: a review*. In *Gesture Workshop*, vol. 1739 of *LNCS*, pp. 103–115, Gif-sur-Yvette, France, Mar. 1999.

[23] YANG, R., S. SARKAR and B. LOEDING: *Enhanced Level Building Algorithm to the Movement Epenthesis Problem in Sign Language*. In *CVPR*, MN, USA, June 2007.

[24] ZAHEDI, M., P. DREUW, D. RYBACH, T. DESELAERS and H. NEY: *Continuous Sign Language Recognition - Approaches from Speech Recognition and Available Data Resources*. In *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pp. 21–24, Genoa, Italy, May 2006.