The SignSpeak Project - Bridging the Gap Between Signers and Speakers

Philippe Dreuw¹, Hermann Ney¹, Gregorio Martinez², Onno Crasborn³, Justus Piater⁴, Jose Miguel Moya⁵, and Mark Wheatley⁶

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.

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 (Crasborn et al., 2007). 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

Signed languages vary like spoken languages do: they are not mutually understandable, and there is typically one or more signed language in every country of the world.

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) is a non-research partner in the SignSpeak project, is a European non-profit making organization which aims to establish and maintain EU level dialogue with the “hearing world” in consultation and cooperation 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, 2008; Wheatley and Pabsch, 2010). 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 (Tervoort, 1953) and
1.3. Research and Challenges in Automatic Sign Language Recognition

In (Ong and Ranganath, 2005) [Y. Wu, 1999] 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 language recognition system.

1.3.1. 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 (von Agris and Kraiss, 2007), or offer small variations only (Dreuw et al., 2008b; Burgeroth et al., 2008), resulting in probably over-fitted language models. Additionally, most self-recorded corpora consists only of a limited number of signers (Vogler and Metaxas, 2001; Bowden et al., 2004).

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 (Farhadi and Forsyth, 2006). In (Buehler et al., 2009; Cooper and Bowden, 2009), 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.

1.3.2. 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.

1.3.3. 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 (Stokoe et al., 1960), 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 (von Agris and Kraiss, 2007), poor scaling (Buehler et al., 2009; Cooper and Bowden, 2009), and unsuitable databases for mostly data-driven approaches (Dreuw et al., 2008b).

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 (Veale et al., 1998) and typically employed rule-based systems, several research groups have recently focussed on data-driven approaches. In (Stein et al., 2006), 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 (Morrissey and Way, 2005) have 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 (Chiu et al., 2007), 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 (San-Segundo et al., 2006), some basic research is done on Spanish and Spanish sign language pair Chinese and Taiwanese sign language. The developments being made in the area of Irish sign language pair English and sign language of the Netherlands with further developments being made in the area of Irish sign language. In (Chiu et al., 2007), 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 (San-Segundo et al., 2006), some basic research is done on Spanish and Spanish sign language.

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 independent, 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). 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).

2.1 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 as explained more detailed in (Dreuw et al., 2008c).

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 – 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 (Vogler and Metaxas, 2001; Yang et al., 2007). 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.

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. The problems described in Section 2. will mainly be tackled by

- analysis of linguistic markers for sub-units and sentence boundaries,
- head and hand tracking of the dominant and non-dominant hand,
- facial expression and body posture analysis,
- analysis of linguistically- and data-driven sub-word units for sign modeling,
- analysis of spatio-temporal across-word modeling,
- signer independent recognition by pronunciation modeling, language model adaptation, and speaker adaptation techniques known from ASR
- contextual and multi-modal translation of sign language by an integration of tracking and recognition features into the translation process

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 sign language related application scenarios would be possible:

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

3.1. Impact on Other Industrial Applications

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.

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 (Verbmobil II) as e.g. presented in (Kanthak et al., 2000), 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 (IWSLT) as e.g. presented in (Mauser et al., 2006).

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 (Forster et al., 2010). The core of the SignSpeak data will stem from the Corpus-NGT. This 72 hour corpus of Sign Language of the Netherlands is the first large open access corpus for sign linguistics in the world. It presently contains recordings from 92 different signers, mirroring both the age variation and the dialect variation present in the Dutch Deaf community (Onno Crasborn and Ros, 2008).

During the SignSpeak project, the existing RWTH-Phoenix corpus (Stein et al., 2006) and Corpus-NGT (Onno Crasborn and Ros, 2008) will be extended to meet these demands (see Table 1). For the SignSpeak project, the limited gloss annotations that were present in the first release of 2008 have been considerably expanded, and sentence-level translations have been added. The smaller RWTH-BOSTON-104 database (Dreuw et al., 2007a) has been extended by a groundtruth annotation for both hand
Table 1: Expected corpus annotation progress of the RWTH-Phoenix and Corpus-NGT corpora.

<table>
<thead>
<tr>
<th></th>
<th>Boston104</th>
<th>Phoenix</th>
<th>Corpus-NGT</th>
<th>Vermobil II</th>
<th>IWSLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>recordings</td>
<td>201</td>
<td>78</td>
<td>400</td>
<td>116</td>
<td>300</td>
</tr>
<tr>
<td>running words</td>
<td>0.8k</td>
<td>10k</td>
<td>50k</td>
<td>30k</td>
<td>80k</td>
</tr>
<tr>
<td>vocabulary size</td>
<td>0.1k</td>
<td>0.6k</td>
<td>2.5k</td>
<td>3k</td>
<td>&gt; 5k</td>
</tr>
<tr>
<td>T/T ratio</td>
<td>8</td>
<td>15</td>
<td>20</td>
<td>10</td>
<td>&lt; 20</td>
</tr>
</tbody>
</table>

Performance

<table>
<thead>
<tr>
<th></th>
<th>11% WER</th>
<th>-</th>
<th>-</th>
<th>15% WER</th>
<th>40% TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opas</td>
<td>Drew et al., 2008a</td>
<td>-</td>
<td>-</td>
<td>(Kanthak et al., 2000)</td>
<td>(Mauser et al., 2006)</td>
</tr>
</tbody>
</table>

and face positions – more than 15,000 frames have been annotated – in order to evaluate the performance of different tracking approaches. Novel facial features (Piater et al., 2010) developed within the SignSpeak project are shown in Figure 3 and will be analyzed for continuous sign language recognition.

For automatic sign language recognition, promising results have been achieved for continuous sign language recognition under lab conditions (von Agris and Kraiss, 2007; Dreuw et al., 2007a). Even if the performances of the automatic learning approaches presented in (Farhadi and Forsyth, 2006) and (Buehler et al., 2009) are still quite low, they represent an interesting approach for further research.

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 (Dreuw et al., 2007a) and translation (Dreuw et al., 2008c) 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 (Dreuw et al., 2008b).

5. Acknowledgements

This work received funding from the European Community’s Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).

6. References


[www-i6.informatik.rwth-aachen.de/~dreuw/database.php]
Figure 3: Facial feature extraction on the Corpus-NGT database (f.l.t.r.): three vertical lines quantify features like left eye aperture, mouth aperture, and right eye aperture; the extraction of these features is based on a fitted face model, where the orientation of this model is shown by three axes on the face: red is X, green is Y, blue is Z, origin is the nose tip.


B. Tervoort. 1953. Structurele analyse van visueel taalgebruik binnen een groep dove kinderen.


