Open Domain Speech Translation: From Seminars and Speeches to Lectures

Christian Fügen[*], *Muntsin Kolss*[*], *Matthias Paulik*[†], *Sebastian Stüker*[*],
Tanja Schultz[†], *Alex Waibel*[*†]

[*]Interactive Systems Labs (ISL)
Universität Karlsruhe (TH), Germany
{fuegen, kolss, stueker, waibel}@ira.uka.de
[†]Interactive Systems Labs (ISL)
Carnegie Mellon University, Pittsburgh, PA, USA
{paulik, tanja}@cs.cmu.edu

ABSTRACT

This paper describes our ongoing work in domain unlimited speech translation. We describe how we developed a lecture translation system by moving from speech translation of European Parliament Plenary Sessions and seminar talks to the open domain of lectures. We started with our speech recognition (ASR) and statistical machine translation (SMT) 2006 evaluation systems developed within the framework of TC-Star (Technology and Corpora for Speech to Speech Translation) and CHIL (Computers in the Human Interaction Loop). The paper presents the speech translation performance of these systems on lectures and gives an overview of our final real-time lecture translation system.

1. INTRODUCTION

Growing international information structures and decreasing travel costs could make the dissemination of knowledge in this globalized world very easy – if only the language barrier could be overcome. Lectures are a very effective method of knowledge dissemination. Such personalized talks are the prefered method since they allow the speakers to tailor their presentation toward a specific audience, and in return allow the listeners to get the most relevant information through interaction with the speaker. In addition, personal communication fosters the exchange of ideas, allows for collaboration, and forms ties between distant units, e.g. scientific laboratories or companies. At the same time it is desirable to allow the presenters of talks and lectures to speak in their native language, since, no matter how proficient in a foreign language, one will always feel more confident in the native tongue. To overcome this obstacle human translators are currently the only solution. Unfortunately, translation service are often prohibitively expensive such that many lectures are not given at all as a result of the language barrier. The use of modern machine translation techniques have the potential to provide translation services at no costs to a wide audience, making it possible to overcome the language barrier and bring the people closer together.
This paper describes our ongoing work in unlimited domain speech translation of lectures starting from systems built within the framework of CHIL and TC-STAR.

CHIL [25], *Computers in the Human Interaction Loop*, aims at making significant advances in the fields of speaker localization and tracking, speech activity detection and distant-talking automatic speech recognition. Therefore, in addition to the near and far-field microphone, seminars were also recorded by calibrated video cameras. The long-term goal is the ability to recognize speech in a real reverberant environment, without any constraint on the number or distribution of microphones in the room nor on the number of sound sources active at the same time.

TC-STAR [20], *Technologies and Corpora for Speech-to-Speech-Translation*, is envisaged as a long-term effort to advance research in all core technologies for Speech-to-Speech Translation (SST) which is a combination of Automatic Speech Recognition (ASR), Spoken Language Translation (SLT) and Text to Speech (TTS). The objective of the project is to make a breakthrough in SST that significantly reduces the gap between human and machine translation performance. The focus hereby is on the development of new algorithms and methods. So far the project targets a selection of unconstrained conversational speech domains – speeches and broadcast news – and three languages: European English, European Spanish, and Mandarin Chinese.

The paper is organized as follows: The developmental work started from our 2006 ASR and SMT evaluation systems for European Parliament Plenary Session (EPPS, TC-STAR) and the NIST Rich Transcription evaluation RT-06S on seminars (CHIL). In Section 3, we first compare the different ASR systems of both domains and show how we merged these systems for lecture recognition. Furthermore, we present first results of acoustic and language model adaptation on the lecture domain. In Section 4, we give statistical machine translation results on text and ASR input for lectures of our 2006 SMT evaluation system for EPPS. In addition, we explain in detail how we adapted our EPPS SMT system towards the more conversational style of lectures and present the corresponding machine translation results. Section 5 provides an overview of our real-time lecture translation system, Section 6 concludes this paper.

2. DEVELOPMENT AND EVALUATION DATA

For the automatic speech recognition and statistical machine translation experiments on lectures, we selected three different lectures as development and evaluation data. The three lectures were given in English by the same non-native speaker on different topics. All lectures were recorded with close talking microphones [3].

Dev: A 24min talk that was held to give a broad overview of current research projects in our lab and is therefore ideal as development set.

t035: A 35min talk held as a conference key-note, only partly covered by the Dev talk, which gives us the opportunity to evaluate how our system behaves on an unseen domain.

t036+: A 31min talk on the same topic as t035, but held in a different environmental setting and situation, which allows us to evaluate the robustness of our system.

For the ASR experiments we used the seminar part of the NIST RT-06S development data and the 2006 EPPS development data as additional data sources.

3. SPEECH RECOGNITION

In this section we first compare the 2006 evaluation systems for European Parliament Plenary Sessions [19] and CHIL seminars [4] and describe the development of a single system, which performs almost as good as the evaluation systems on both domains respectively. This is followed by the presentation of the system's performance on the lecture domain. Lectures are an ideal showcase for speaker and domain adaptation tasks, since the lecturer and the topic might be known in advance [1]. Therefore, we describe acoustic and language model adaptation results in the last part of this section. Different from the work [3] we will in this paper take the 2006 EPPS evaluation into consideration for the development of our lecture recognition system.

All speech recognition experiments were done using the Janus Recognition Toolkit (JRTk) featuring the Ibis decoder [17]. For language modeling, we used the SRI Language Modeling Toolkit (SRILM) [18].

3.1. Data

For acoustic model training, we selected the following corpora: ICSI and NIST meeting recordings [9, 12], TED lectures [11], CHIL seminars [25], and European Parliament Plenary Sessions (EPPS) [8]. Because of the results given in [4] we have neither used the ISL meeting corpus nor the Hub4 Broadcast News corpus due to their channel mismatch: both corpora were recorded with lapel microphones. Table 1 gives an overview of the total amount of speech in the different corpora. More information about the respective corpora can be found in the cited literature.

	ICSI	NIST	TED	CHIL	EPPS
speakers	463	77	52	67	1894
duration	72h	13h	13h	10h	80h

Table 1: Number of speakers and total amount of speech data used for acoustic model training.

For language model training, some additional text data was used on top of the 2006 evaluation systems' [4, 19] language model training data. Altogether, the following corpora were available: Talks, text documents from TC-STAR and CHIL, EPPS transcripts, EPPS final text editions, AMI meeting data, non-AMI meeting data (ISL, ICSI, NIST), TED lectures, CHIL seminars, broadcast news data, UN (United Nations) text data released by ELDA, recent proceedings data (2002 - 2005), web data from UWash (related to ISL, ICSI, and NIST meetings) and web data collected56964296(t)-227.819(()-4.2603(S)1.92982(R)4.52619(I
ne in(n)-24908]TJ T*[(c)-1.66516(781(U)-0.700861(N)-470865(i)0.965521(
o521(o)18.20638(n)-5.892(e(6..471.63(i)0.96522(e)-1.67128(m)-4.92563(s)3
-21(n)-5.8963(i)0.965521(s)-237.38(r)-4.259123(t)0.963071(o)-367(n)-5.891

	talks	docs	eppsS	eppsT	nAMI	AMI	TED	CHIL	BN	UN	proc	UWash	wCHIL
words	93k	192k	750k	33M	1.1M	200k	98k	45k	131M	42M	23M	147M	146M
EPPS			35%	54%					9%	2%			
CHIL					15%	8%	0.6%	25%	0.8%		24%	12%	15%
Dev	36%	1%		12%			3%		8%		9%	11%	19%

Table 2: Amount of language model training data in words together with their interpolation weights for the different domains. 'Dev' is the lecture development set as described in Section 2. Empty cells indicate that the data was not useful for that domain.

We used a three pass decoding setup. As in [4], the first pass uses incremental speaker based vocal tract length normalization (VTLN) and constrained MLLR estimation and is decoded with the semi continuous models (4) using tight search beams. The second pass uses the same semi continuous acoustic models as pass one, but before decoding, MLLR [13] adaptation together with an estimation of fixed VTLN and constrained MLLR parameters is performed. For this, the confidence weighted hypomal(e)-1.66393(s)3.566516(k)6.18644(-1.66393(t)0.11.0398 TL T*[(o)-5.89054(f)-365 used exactly the same decoding dictionaries and language models as for the EPPS and RT-06S evaluation systems.

CHIL Seminars For the CHIL seminars we used the same language models and dictionaries as described in [4]. T-1.6658993(e)-362.937(l)0.965521(a)-1.66516(n)-5.88993(g)-5.89054(u)-5.89054(a)-1.66516(g)-5.88993(e)-350.923(m)-4.9244(o)-5.889 meetings, TED, some CHIL data, BN, proceedings and web data related to meetings and CHIL lectures. The interpolation weights, which were tuned on held-out CHIL data are shown in Table 2. The language model has a perplexity of 130 on the RT-06S development data, while 16% 4-grams, 41% 3-grams, 39% 2-grams, and 4% 1-grams were used. The dictionary consists of about 59k pronunciation variants defined over a vocabulary of 52k. It has an out-of-vocabulary (OOV) rate of 0.65 on the RT-06S development data.

As can be seen in table 3 for the above described different system passes, acoustic models trained on EPPS alone or additionally including TED (TED+EPPS) is significant worse than the other two systems. The performance of the two other systems is nearly identical, which means that adding the EPPS data to the acoustic model training data used in RT-06 (ICSI+NIST+TED) dol(e)-1.66393(s)-273.429(n)-5.8887(o)-5.869(m)-4-276.031(h)-5.8887(u)-5.89115(r)-4.25908(-4-263.9 (ICSI+NIST+TED) performs worst. Fur-40.970422(h)-5.89605(e)-1.66638(r)-4.25908(m)-4.92563(o)-5.89115(r)-4.25908(e)-1.66638(,)- a corpus containing European English as well. By addingm]TJ 11.0398 TL T*[(t)0.963071(h)-5.8887(e)-254.586(m)-4.92563(e)-1.66393(gives the same resumts compared to the TED+EPPS system.

	1st	2nd

3.4. Lecture Domain

Based on the perplexities and OOV-rates on Dev shown in Table 5 we selected the language model and dictionary built for the CHIL seminars for our baseline experiments. Not surprisingly, this selection holds also for the evaluation talks. The EPPS language model and vocabulary is, due to the large amount of in-domain data, too specific. The OOV-rates of the RT-06S (CHIL) vocabulary and for t036+ are surprisingly low – the only explanation for that is this talk is not very specific.

	Dev		t035		t036+	
	PPL	OOV	PPL	OOV	PPL	OOV
CHIL	173	0.22	117	0.27	186	0.09
EPPS	205	1.29	230	1.83	229	1.72

Table 5: Perplexities (PPL) and OOV-rates of the CHIL and EPPS language models and vocabularies.

As can be seen in Table 6, the acoustic model trained on all data performs significantly better than the other models. For this reason we selected this model for our further experiments. The baseline results on the lecture evaluation talks are shown in Table 7. With the training setup developed for RT-06S we significantly improved our results compared to the acoustic models developed in [3] (MS11 column in Table 7). Furthermore, it can be seen that the system performs quite well on unseen domains (t035) and different environments (t036+).

Model Adaptation Experiments The baseline experiments were performed with unsupervised adaptation. As mentioned above, for lectures, speaker and topic are often known in advance. Therefore, the lecture domain is ideal for applying supervised acoustic and language model adaptation. As will be shown, this allows us to reduce the decoding setup from three to only one single decoding pass without any loss in performance and is the first step towards a real-time lecture translation system.

For acoustic model adaptation an additional amount of around 7 hours of speech for the same speaker was available. For the adaptation experiments subsets of this data with different lengths were used to compute VTLN and constrained MLLR (FSA) parameters and to perform model based MLLR adaptation. The results can be seen in Table 8. While the adaptation works quite well on the evaluation talks – the 7hrs results are similar to those achieved after CNC with the baseline systems – the results on the

	1st	2nd	3rd	cnc
EPPS	23.9	–.-	–.-	–.-
TED+EPPS	23.4	–.-	–.-	–.-
ICSI+NIST+TED+EPPS	21.4	16.2	15.0	15.5
ICSI+NIST+TED	24.3	–.-	–.-	–.-

Table 6: Baseline results on Dev with the CHIL dictionary and language model. The CHIL data was used in all systems for acoustic model training.

	1st	2nd	3rd	cnc	MS11
t035	17.3	12.6	12.1	12.2	12.7
t036+	16.7	12.0	11.6	11.5	12.4

Table 7: Baseline results on the evaluation talks t035 and t036+. The MS11 column contains the final (CNC) results with the acoustic model trained in [3].

	0.5hrs	1.5hrs	3.5hrs	7hrs	sup
Dev	20.9	20.0	19.5	18.9	12.0
t035	14.2	13.1	12.6	12.1	10.1
t036+	13.3	12.3	11.5	10.7	9.3

Table 8: Acoustic model adaptation results with different amounts of adaptation data. In the column 'sup', supervised adaptation was performed on the particular talk itself.

Dev talk are significantly worse. This is due to a large channel mismatch between the adaptation material and the Dev talk. To confirm this, we adapted on the particular talk itself and achieved reasonable results for all talks (see column 'sup' in Table 8). It can also be seen, that doubling

4.1. Phrase Alignment

To find a translation for a source phrase $\tilde{f} = f_1 ... f_l$ we restrict the general word alignment: Words inside the source phrase align to words inside the target phrase, and words outside the source phrase align to words outside the target phrase. This constrained alignment probability is calculated using the well-known IBM1 word alignment model, but the summation of the target words is restricted to the appropriate regions in the target sentence. Also, the position alignment probabilities are adjusted accordingly [23]. Optimization is over the target side boundaries i_1 and i_2.

$$\begin{aligned} p_{i_1,i_2}(f|e) = & \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \\ & \times \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{1}{k} p(f_j|e_i) \\ & \times \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1..i_2)} \frac{1}{I-k} p(f_j|e_i) \end{aligned} \quad (1)$$

Similar to $p_{i_1,i_2}(f|e)$ we can calculate $p_{i_1,i_2}(e|f)$, now summing over the source words and multiplying along the target words.

To find the optimal target phrase we interpolate the log probabilities and take the pair (i_1, i_2) that gives the highest probability. The interpolation factor c can be estimated on a development test set.

The scores calculated in the phrase alignment are alignment scores for the entire sentence. As phrase translation probabilities we use the second term in Eqn. 1.

4.2. Decoder

The beam search decoder combines all model scores to find the best translation. In these experiments, the different models used were: (1) The translation model, i.e. the word-to-word and phrase-to-phrase translations extracted from the bilingual corpus according to the new alignment method described in this paper. (2) A trigram language model. The SRI language model toolkit was used to train the models [18]. (3) A word reordering model, which assigns higher costs to longer distance reordering. We use the jump probabilities $p(j|j')$ of the HMM word alignment model [24] where j is the current position in the source sentence and j' is the previous position. (4) Simple word and phrase count models. The former is essentially used to compensate for the tendency of the language model to prefer shorter translations, while the latter can be used to give preference to longer phrases. For each model a scaling factor can be used to modify the contribution of this model to the overall score.

The decoding process is organized into two stages: First, the word-to-word and phrase-to-phrase translations and, if available, other specific information like named entity translation tables are inserted into a translation lattice. In the second step, we find the best combinations of these partial translations, such that every word in the source sentence is covered exactly once. This amounts to doing a best path search through the translation lattice, which is extended to allow for word reordering: Decoding proceeds essentially along the source sentence. At each step, however, the next word or phrase to be translated may be selected from all words laying or phrases starting within a given look-ahead window from the current position [22].

4.3. Training Data

For training the baseline translation systems, the parallel EPPS corpus was used. For English-Spanish, a version was created by RWTH Aachen within TC-STAR [8]. The English-to-German models were trained on the EPPS data as provided by Philipp Koehn [10].

In addition, a small number of lectures similar in style to our development and evaluation data was collected, transcribed, and translated into Spanish and German. Altogether, parallel lecture corpora of about 12,000 words were available in each language.

4.4. Model Adaptation

Adapting the MT component of our EPPS translation system towards the more conversational style of lectures was accomplished by a higher weighting of the available lecture data in two different ways. First, for computing the translation models, the small lecture corpora were multiplied several times and added to the original EPPS training data. This yielded a small increase in MT scores.

Secondly, for (target) language model computation, a small tri-gram LM was computed on t035 and then interpolated with the original EPPS language model, whereas the interpolation weight was chosen in order to minimize the perplexity on the development set. In this manner the perplexity on the Dev talk could be reduced from 645 to 394 for German and from 543 to 403 for Spanish. To further adapt the target language models, we collected Spanish and German web data with the help of tools provided by the University of Washington [21]. A small amount of the used search queries were hand written, however, most search queries were automatically created by using the most frequent tri-grams found in the Dev talk. Approximately $1/4$ of all development set tri-grams were used for this. The German and Spanish web corpora collected in this manner consisted of 175M words and 120M words, respectively. The web corpora were again added to the existing LMs by interpolation, which yielded a perplexity of 200 for German and 134 for Spanish. The corresponding perplexities on the t036+ talks are 617 and 227, respectively.

The effects of translation model and language model adaptation, as well as the results of the final system, combining both adaptation steps, are shown in tables 10 and 11 for English-to-Spanish and English-to-German, respectively.

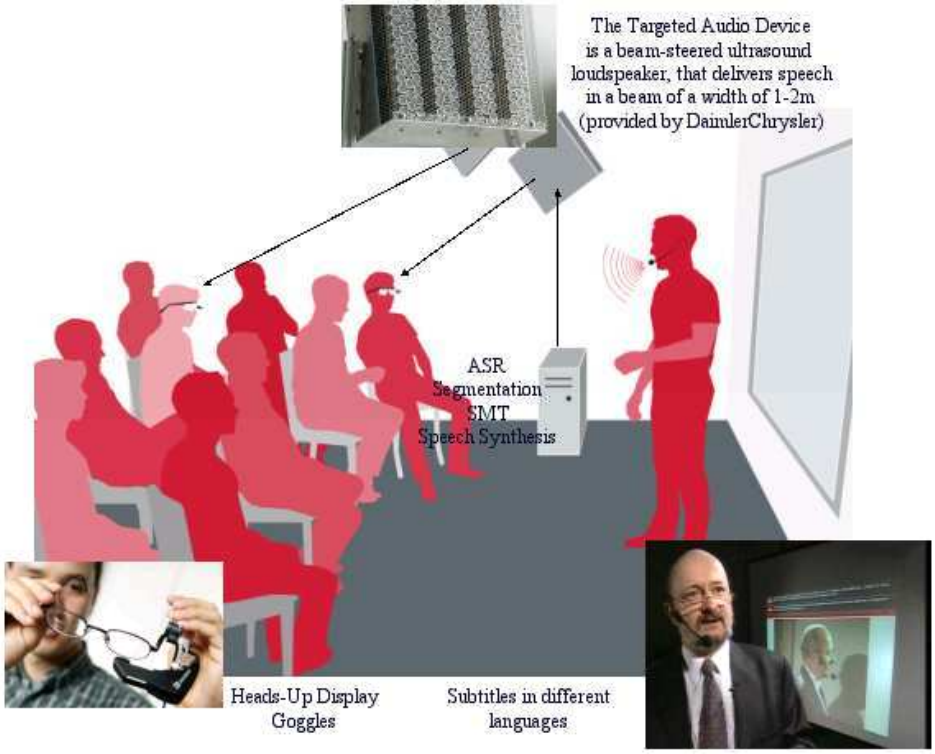

Figure 1: The lecture transcription system.

STAR – Technology and Corpora for Speech to Speech Translation – (Grant number IST-506738). The authors would like to thank Susanne Burger, Maria Kernecker, Tim Notari, and Silja Hildebrand for transcribing and translating the lecture development and evaluation data and Florian Kraft for providing the language model data used for the RT-06S evaluation system.

BIBLIOGRAPHIE

[1] M. Cettolo, F. Brugnara, and M. Federico. Advances in the Automatic Transcription of Lectures. In *ICASSP*, Montreal, Canada, 2004.

[2] K. Linhard D. Olszewski, F. Prasetyo. Steerable Highly Directional Audio Beam Louspeaker. In *Proc. of the Interspeech*, Lisboa, Portugal, September 2006.

[3] C. Fügen, M. Kolss, D. Bernreuther, M. Paulik, S. Stüker, S. Vogel, and A. Waibel. Open Domain Speech Recognition & Translation: Lectures and Speeches. In *ICASSP*, Toulouse, France, 2006.

[4] C. Fügen, M. Wölfel, J. W. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani. Advances in Lecture Recognition: The ISL RT-06S Evaluation System. In *submitted to Interspeech 2006*, Pittsburgh, PA, USA, September 2006.

[5] Christian Fügen, Martin Westphal, Mike Schneider, Tanja Schultz, and Alex Waibel. LingWear: A Mobile Tourist Information System. In *Proc. of the Human Language Technology Conf. (HLT)*, San Diego, California, March 2001. NIST.

[6] M. J. F. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical report, Cambridge University, Cambridge, United Kingdom, 1997.

[7] M. J. F. Gales. Semi-tied covariance matrices. In *ICASSP*, 1998.

[8] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney. Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus. *ICASSP*, 2005.

[9] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. The ICSI Meeting Project: Ressources and Research. In *Proc. of the ICASSP Meeting Recognition Workshop*, Montreal, Canada, May 2004. NIST.

[10] P. Koehn. Europarl: A Multilingual Corpus for Evaluation of Machine Translation, 2003. http://people.csail.mit.edu/koehn/publications/europarl.

[11] L.F. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann. The Translanguage English Database TED. In *ICSLP*, volume LDC2002S04, Yokohama, September 1994. LDC.

[12] Linguistic Data Consortium (LDC). ICSI, ISL and NIST Meeting Speech Coprora at LDC, 2004. http://www.ldc.upenn.edu catalog IDs LDC2004S02, LDC2004S05, LDC2004S09.

[13] C. J. Leggetter and P. C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185, 1995.

[14] L. Mangu, E. Brill, and A. Stolcke. Finding Consensus among Words: Lattice-based Word Error Minimization. In *EUROSPEECH*, 1999.

[15] NIST. NIST MT evaluation kit version 11a, 2004. http://www.nist.gov/speech/tests/mt.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center, 2002.

[17] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment. In *ASRU*, Trento, Italy, 2001.

[18] A. Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *ICSLP*, Denver, Colorado, USA, 2002.

[19] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, F. Kraft Q. Jin, M. Paulik, M. Raab, Y.-C. Tam, and M. Wölfel. The ISL TC-STAR Spring 2006 ASR Evaluation Systems. In *submitted to the TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, June 2006.

[20] TC-Star. Technology and corpora for speech to speech translation, 2004. http://www.tc-star.org.

[21] UWash. University of Washington, web data collection scripts, 2006. http://ssli.ee.washington.edu/projects/ears/WebData/web_data_collection.html.

[22] S. Vogel. SMT Decoder Dissected: Word Reordering. In *Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Beijing, China, 2003.

[23] S. Vogel. PESA: Phrase Pair Extraction as Sentence Splitting. In *Machine Translation Summit 2005*, Thailand, 2005.

[24] S. Vogel, H. Ney, and C. Tillmann. HMM-based Word Alignment in Statistical Translation. In *COLING 96*, pages 836–841, Copenhagen, 1996.

[25] A. Waibel, H. Steusloff, and R. Stiefelhagen. CHIL – Computers in the Human Interaction Loop. In *5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, April 2004. http://chil.server.de.