Open Domain Speech Translation: From Seminars and Speeches to Lectures

Christian Fügen[*], **Muntsin Kolss**[*], **Matthias Paulik**[†], **Alex Waibel**[*†]

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

Abstract

This paper describes our ongoing work in open domain 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 and statistical machine translation 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 preferred 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 services 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 (Waibel et al., 2004), *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. 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 (TC-Star, 2004), *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) (Stüker et al., 2006) and the NIST Rich Transcription evaluation RT-06S on seminars (CHIL) (Fügen et al., 2006b). 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 (ASR) and statistical machine translation (SMT) experiments on lectures, we selected three different lectures as development and evaluation data. These lectures were held in non-native English by the same speaker on different topics and were recorded with close talking microphones (Fügen et al., 2006a).

Dev: This 24min talk was held to give a broad overview of current research projects in our lab.

t035: A 35min talk held as a conference key-note, which is only partly covered by the Dev talk.

t036+: A 31min talk on the same topic as t035, but held in a different environment and situation.

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 (Stüker et al., 2006) and CHIL seminars (Fügen et al., 2006b) and show how we developed a single system, which performs almost as good as the evaluation systems on both domains. This is followed by the presentation of the system's performance on the lecture domain.

All speech recognition experiments were done using the Janus Recognition Toolkit (JRTk) featuring the Ibis decoder (Soltau et al., 2001). For language modeling, we used the SRI Language Modeling Toolkit (SRILM) (Stolcke, 2002).

3.1. Data

For acoustic model training, we selected the following corpora: ICSI and NIST meeting recordings (Janin et al., 2004; (LDC), 2004), TED lectures (Lamel et al., 1994), CHIL seminars (Waibel et al., 2004), and European Parliament Plenary Sessions (EPPS) (Gollan et al., 2005). Given the results in (Fügen et al., 2006b) we have not used the ISL meeting corpus as well as 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.

	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 for the acoustic model training data.

For language model training, some additional text data was used on top of the 2006 evaluation systems' (Fügen et al., 2006b; Stüker et al., 2006) language model training data. Altogether, the following corpora were available: Talks, text documents from TC-STAR and CHIL, EPPS transcripts, EPPS final text editions, non AMI meeting data (ISL, ICSI, NIST), AMI meeting data, 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 collected for RT-06S (related to CHIL seminars). Table 2 shows the amount of words available for each corpus.

3.2. System Description

The acoustic models used in the experiments below were all trained in the same way, resulting in a size of 16,000 distributions over 4,000 models, with a maximum of 64 Gaussians per model. These models were all based on the same quint-phone context decision tree and phoneme set that was used for the RT-06S evaluation system. Furthermore, also the acoustic model training setup was taken from the RT-06S system: (1) a first incremental growing of Gaussians, (2) estimation of the global STC transform (Gales, 1998), (3) a second incremental growing of Gaussians. To train the distributions for the semi-continuous system and to compensate for the occasionally worse fixed-state alignments, 2 iterations of Viterbi training were performed (4). For the SAT models, 4 additional iterations of SAT

Viterbi training by using constrained MLLR in the feature space (FSA) (Gales, 1997) were applied (5). We use an MFCC FFT front-end with a 42-dimensional feature space after linear discriminant analysis (LDA) and a global STC transform with utterance-based cepstral mean subtraction (CMS). More details can be found in (Fügen et al., 2006b).

3.3. Baseline Experiments and Comparisons

The goal was to build a single acoustic model for both domains, EPPS and CHIL seminars and to finally use this acoustic model on the lecture data. For this, we compared different acoustic models trained on different subsets of the acoustic training material described in Section 3.1.. All subsets contain the CHIL corpus, which is therefore not explicitly mentioned in the table rows below.

We used a three pass decoding setup to be comparable with other evaluation systems. 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 (Leggetter and Woodland, 1995) adaptation together with an estimation of fixed VTLN and constrained MLLR parameters is performed. For this, the confidence weighted hypotheses of the previous pass are being used. For the third pass, the FSA-SAT acoustic models (5) are used together with the same adaptation scheme applied in pass two. After that, confusion network combination (CNC) (Mangu et al., 1999) is being performed, using the lattices of the third pass only. We used exactly the same decoding dictionaries and language models as for the EPPS and RT-06S evaluation systems.

3.3.1. CHIL Seminars

For the CHIL seminars we used the same language models and dictionaries as described in (Fügen et al., 2006b). The 4-gram language model was trained on AMI and non-AMI 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 around 59k pronunciation variants over a vocabulary of 52k. It has an OOV-Rate of 0.65 on the RT-06S development data.

As can be seen in Table 3, acoustic models trained on EPPS alone or additionally including TED are 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) does not hurt (but also does not improve the overall results).

3.3.2. European Parliament Plenary Sessions

For the European Parliament Plenary Sessions we used the language models and dictionaries as described in (Stüker et al., 2006). The 4-gram language model was trained on EPPS transcriptions and final text editions, BN, and UN and reached a perplexity of 93 on the 2006 EPPS development data, whereas 29% 4-grams, 36% 3-grams, 32% 2-grams,

	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: 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.

CHIL	1st	2nd	3rd	cnc
EPPS	40.3	–.-	–.-	–.-
TED+EPPS	38.7	–.-	–.-	–.-
ICSI+NIST+TED+EPPS	34.1	27.5	26.2	25.5
ICSI+NIST+TED	34.0	27.1	26.0	25.5

Table 3: Results on the RT06 development data. The CHIL data was used in all systems for AM training.

and 4% 1-grams were used. The interpolation weights were tuned on the 2005 EPPS development data and are shown in Table 2. The dictionary for EPPS consists of 45k pronunciations over a vocabulary of 40k and has an OOV-Rate of 0.43 on the 2006 EPPS development data.

As can be seen in Table 4 the last system trained without EPPS performs worst. Furthermore, compared to the acoustic model used for the 2006 EPPS evaluation (MS23, (Stüker et al., 2006)), the acoustic model training setup developed for RT-06S is significantly better (MS23 vs. EPPS rows). An additional gain can be seen by adding TED, which is also a corpus containing European English. By adding the meeting data, the system improves not further, instead it is in between the EPPS and TED+EPPS systems. Nevertheless, after doing confusion network combination, it performs identical to the TED+EPPS system.

	1st	2nd	3rd	cnc
MS23	22.6	–.-	–.-	–.-
EPPS	20.8	15.4	14.7	14.5
TED+EPPS	20.1	14.8	14.3	14.1
ICSI+NIST+TED+EPPS	20.6	15.1	14.6	14.1
ICSI+NIST+TED	29.1	–.-	–.-	–.-

Table 4: Results on the 2006 EPPS development data. The CHIL data was used in all systems for AM training, besides MS23. MS23 specifies the 2006 EPPS evaluation setup.

Compared to the CHIL seminars, the EPPS results are much better. The reason for that lies in the huge available amount of acoustic and language model in-domain training data for EPPS compared to CHIL, where only a very small amount of in-domain data is available. Furthermore, the language used in the European Parliament is more formal and therefore less spontaneous. This leads also to a better OOV-rate and language model perplexity with a higher n-gram coverage for larger n-grams.

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, that 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 (Fügen et al., 2006a) (MS11 column in Table 7). Furthermore, it can be seen that the system performs quite well on unseen domains (t035) and different environments (t036+).

3.4.1. Model Adaptation Experiments

Lectures are an ideal domain for doing adaptation, because the lecturer and also the topic might be known in advance (Cettolo et al., 2004). Therefore, we give first acoustic and language model adaptation results. As can be seen, this allows us to reduce the decoding setup from three to only a single decoding pass without any loss in WER and is the first step towards a real-time lecture recognition and 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 durations 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

	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 (Fügen et al., 2006a).

	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.

CNC with the baseline systems – the results on the 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 reached for all talks reasonable results (see column sup in Table 8). It can also be seen, that doubling the adaptation data results in a relative gain of around 0.5% in WER.

For language model adaptation we only did a first experiment by tuning the interpolation weights and reselecting the different corpora with respect to the lecture domain. The interpolation weights, tuned on some held-out data and the selected corpora can be seen in Table 2. Thereby the perplexity on the Dev talk could only be reduced slightly from 173 to 168. Nevertheless we saw significant gains in WER on all lectures, which are reported in Table 9.

4. Statistical Machine Translation

In this section, we describe the statistical machine translation (SMT) component in our lecture translator that was used to translate the lectures in Section 2. from English to Spanish and German. The underlying phrase-based SMT system was originally developed within TC-STAR for translating speeches from the European Parliament Plenary Sessions (EPPS). In these experiments, we used loose coupling, passing the first-best hypothesis from the recognizer to the translation component. Translation results are reported using the well known evaluation metrics BLEU (Papineni et al., 2002) and NIST (NIST, 2004). All MT scores were calculated using case-insensitive scoring and one reference translation per test set.

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

	unadapted	adapted	PPL
Dev	18.9	16.1	168
t035	12.1	10.5	165
t036+	10.7	9.1	193

Table 9: Language model adaptation results on top of the acoustic model adaptation on 7hrs of speech. Perplexities should be compared with Table 5.

outside the source phrase align 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 (Vogel, 2005). Optimization is over the target side boundaries i_1 and i_2.

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

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 (Stolcke, 2002). (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 (Vogel et al., 1996) 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 (Vogel, 2003).

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 (Gollan et al., 2005). The English-to-German models were trained

system	NIST	Bleu
baseline (EPPS)	4.71 (5.61)	15.41 (20.54)
TM-adaptation	4.78 (5.67)	16.05 (21.43)
LM-adaptation	5.10 (5.99)	17.58 (22.90)
final system	5.22 (6.11)	18.57 (24.00)

Table 10: English-Spanish lecture translation system on t036+. Results on manual transcripts are shown in brackets.

system	NIST	Bleu
baseline (EPPS)	4.00 (4.71)	09.32 (12.53)
TM-adaptation	4.29 (5.06)	11.01 (14.95)
LM-adaptation	4.37 (5.12)	11.67 (14.96)
final system	4.67 (5.47)	13.22 (17.25)

Table 11: English-German lecture translation system on t036+. Results on manual transcripts are shown in brackets.

on the EPPS data as provided by Philipp Koehn (Koehn, 2003). 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 (UWash, 2006). 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 out 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. The significantly lower MT scores for the English-to-German translation direction are mostly due to long distance dependencies and compound words which are inherent to the German language. In absolute terms, the translation performance on this difficult task is still quite poor when compared with tasks for which large amounts of training data similar in style is available, such as the TC-STAR EPPS task. Nevertheless, small amounts of lecture data were sufficient to significantly improve performance, especially when amplified by using language model adaptation with similar web data.

5. Real-time Lecture Translation System

For our current version of a real-time lecture translation system, which simultaneously translates lectures given in English into Spanish and German, we integrated the above described speech recognition and machine translation systems together with a sentence segmentation component and a speech synthesis into a client-server framework similar to the one described in (Fügen et al., 2001).

To reach real-time end-to-end performance, we had to tune the above described single pass speech recognizer to run faster than real-time, by further restricting the beam search, which resulted in an increase in WER to about 13% on the evaluation talks. The other system components did not need further tuning.

To keep the latency of the system as short as possible, the speech recognizer already starts to decode, while the speaker is talking and continously returns partial back traces with first best hypotheses. Since the machine translation awaits complete sentences as input, we merged the partial hypotheses together, and resegmented them to sentence like segments. This means, that different from other speech transcription systems no speech segmentation was performed before processing it by the speech recognizer, instead it was done afterwards, to have the ability to tune the segmentation boundaries with respect to optimal machine translation performance. Currently, the segmentation is done at silence regions only, whereby additional thresholds are defined to produce segments with a length of about five to ten words. Thereby, the latency of the system could be limited to a maximum of about five seconds. We plan for more sophisticated segmentation algorithms in the future.

An overview of the real-time lecture translation system is given in Figure 1. As can be seen, the system can deliver the output in different ways:

Subtitles: Simultaneous translations can be projected to the wall as subtitles. This is suitable if the number of output languages is small.

Heads-Up Display Goggles: When there is not enough space on a wall or canvas, heads-up display goggles can be worn to see the simultaneous translation as subtitles. Furthermore, other participants are not disturbed by the subtitles.

Targeted Audio Another solution for providing the simultaneous translation without disturbing others is the so-called targeted audio device (Olszewski et al., 2006). The targeted audio device is a beam-steered loudspeaker, consisting of several small ultrasound loudspeakers. It outputs audio in a beam with a width of about 1-2 meters. People sitting within the beam are able to hear everything, people outside the beam do not. In future applications, several such targeted audio devices could be assigned in various languages to accommodate each participant in the lecture room.
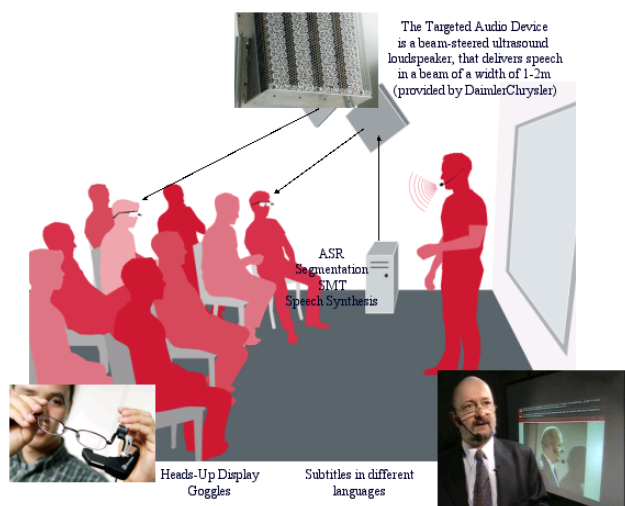
The Targeted Audio Device is a beam-steered ultrasound loudspeaker, that delivers speech in a beam of a width of 1-2m (provided by DaimlerChrysler)

ASR
Segmentation
SMT
Speech Synthesis

Heads-Up Display Goggles

Subtitles in different languages

Figure 1: The lecture transcription system.

6. Conclusion

In this paper, we presented our work in taking first steps towards building open domain speech translation systems. We have successfully developed an ASR system for lectures by merging the evaluation systems for European Parliament Plenary Sessions and CHIL seminars. Furthermore, we combined the resulting system with the translation system used in TC-STAR to translate lectures on a new domain from English to Spanish and German. The ASR system performance exceeds our expectations, demonstrating the feasibility of designing open domain recognition systems. For translation, lectures still pose a significant challenge. Nevertheless, small amounts of lecture data were sufficient to significantly improve performance, especially when amplified by using language model adaptation with similar web data.

7. Acknowledgments

8. References

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

C. Fügen, M. Westphal, M. Schneider, T. Schultz, and A. Waibel. 2001. LingWear: A Mobile Tourist Information System. In *Proc. of the Human Language Technology Conf. (HLT)*, San Diego, California, March. NIST.

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

C. Fügen, M. Wölfel, J. W. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani.

2006b. Advances in Lecture Recognition: The ISL RT-06S Evaluation System. In *To be published at Interspeech 2006*, Pittsburgh, PA, USA, September.

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

M. J. F. Gales. 1998. Semi-tied covariance matrices. In *ICASSP*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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