Mach Translat (2007) 21:209–252
DOI 10.1007/s10590-008-9047-0

Simultaneous translation of lectures and speeches

Christian Fügen · Alex Waibel · Muntsin Kolss

Received: 27 August 2008 / Accepted: 4 November 2008 / Published online: 22 November 2008

Abstract With increasing globalization, communication across language and cultural boundaries is becoming an essential requirement of doing business, delivering education, and providing public services. Due to the considerable cost of human translation services, only a small fraction of text documents and an even smaller percentage of spoken encounters, such as international meetings and conferences, are translated, with most resorting to the use of a common language (e.g. English) or not taking place at all. Technology may provide a potentially revolutionary way out if real-time, domain-independent, simultaneous speech translation can be realized. In this paper, we present a simultaneous speech translation system based on statistical recognition and translation technology. We discuss the technology, various system improvements and propose mechanisms for user-friendly delivery of the result. Over extensive component and end-to-end system evaluations and comparisons with human translation performance, we conclude that machines can already deliver comprehensible simultaneous translation output. Moreover, while machine performance is affected by recognition errors (and thus can be improved), human performance is limited by the cognitive challenge of performing the task in real time.

C. Fügen · A. Waibel · M. Kolss
International Center for Advanced Communication Technologies (InterACT), Fakultät für Informatik,
Universität Karlsruhe (TH), Am Fasanengarten 5, 76131 Karlsruhe, Germany

C. Fügen
e-mail: fuegen@ira.uka.de

A. Waibel (✉)
International Center for Advanced Communication Technologies (InterACT), School of Computer
Science, Carnegie Mellon University, 407 S. Craig Street, Pittsburgh, PA 15213, USA
e-mail: ahw@cs.cmu.edu

C. Fügen · A. Waibel
Mobile Technologies LLC, 619 Windsor Avenue, Pittsburgh, PA 15221-4337, USA

Keywords Simultaneous translation · Interpretation · Speech-to-speech translation · Spoken language translation · Machine translation · Speech recognition · Lecture recognition · Lectures · Speeches

Abbreviations

AMI	Meeting transcription data from Augmented Multi-party Interaction (see Table 3)
ASR	Automatic speech recognition
BN	Data from broadcast news corpora (see Table 3)
CHIL	Computers in the human interaction loop
cMLLR	Constrained maximum likelihood linear regression
DG	Directorate general
EC	European Commission
EM	Expectation maximization
EPPS	European Parliament plenary sessions
GALE	Global Autonomous Language Exploitation
GWRD	Data from Gigaword corpus (see Table 3)
HNSRD	Data from UK Parliament debates (see Table 3)
ICSI	(Data recorded at) International Computer Science Institute
JRTk	Janus Recognition Toolkit
MLLR	Maximum likelihood linear regression
MT	Machine translation
MTG	Meeting transcription data (see Table 3)
NIST	National Institute of Standards and Technology—MT evaluation measure (Doddington 2002) Data recorded at NIST
OOV	Out of vocabulary
PROC	Data from conference proceedings (see Table 3)
RTF	Real-time factor
RT-06S	NIST 2006 Rich Transcription evaluation
RWTH	Rheinisch-Westfälische Technische Hochschule (Aachen)
SMNR	Lectures and seminars from the CHIL project
SMT	Statistical machine translation
SRI	Stanford Research Institute
SST	Speech-to-speech translation
STR-DUST	Speech TRanslation: Domain-Unlimited, Spontaneous and Trainable
SWB	Data from switchboard transcriptions (see Table 3)
TC-STAR	Technologies and Corpora for Speech-to-Speech-Translation
TED	Translanguage English database
TH	Technische Hochschule
TTS	Text to speech
UKA-...	See Table 3
UW-M	See Table 3
VAD	Voice activity detection
VTLN	Vocal tract length normalization
WER	Word error rate

1 Introduction

With advancing globalization, effective cross-cultural communication is rapidly becoming an essential part of modern life. Yet, an estimated 4,000 to 6,000 languages still separate the people of the world, and only a small intellectual elite gather by agreeing to communicate in a common language, usually English. This situation creates a dilemma between a need to integrate and interact and the need to maintain cultural diversity and integrity. It also limits international discourse to an adopted communication medium that does not always reflect the culturally nuanced, individual subtleties of the global human experience. To provide access to other languages unimpeded, however, requires translation, but due to the enormous cost of human translation, only a small fraction of text documents are presently translated and only a handful of human spoken encounters are actually interpreted, if they even take place at all as a result of the separation.

All this could change if affordable, real-time, simultaneous speech-to-speech translation (SST) was possible by machine. Machines never suffer fatigue, nor are they limited in principle by memory size or processing speed. They could potentially also open communication channels between uncommon spoken languages, so as to broaden the exchange and interaction between the people of the world.

As a result, automatic SST as a research endeavor has been enjoying increasing attention. After humble first beginnings in the 1990s, when limited phrase-based and first domain-dependent speech translation systems were proposed (Waibel et al. 1991; Morimoto et al. 1993) large-scale domain-unlimited translation systems are now beginning to emerge (Waibel and Fügen 2008). Three recent projects in particular have begun to advance the state of the art from domain-dependent to domain-independent capabilities: the NSF-ITR project STR-DUST, the EC-IP project TC-STAR, and the DARPA project GALE (see Sect. 3 for more details about these projects). They bring within reach the translation of broadcast news and political speeches between several major languages.

As systems developed within these three projects focus on the offline processing of huge data material, translation quality instead of speed was of primary importance. As we will show in this paper, however, even simultaneous domain-unlimited translation is possible thereby opening up the range of applications to interactive lectures, seminars, speeches, and meetings. To do so, we have combined advances in speech translation technology from these projects with advances in human interface technology developed under the European Commission (EC) integrated project CHIL (Computers in the Human Interaction Loop) (Waibel et al. 2004; Waibel and Stiefelhagen 2009).

While the work within CHIL focused on the delivery and interface aspect of a simultaneous translation service and the integration into the Smart Room, core recognition and translation components were advanced under TC-STAR. In the present paper, we describe how these technologies were extended to deliver real-time simultaneous speech translation of spontaneous lectures. We will discuss how the added requirements found in technical university lectures and seminars, such as real-time performance, online sentence-level segmentation, special terms for lecture vocabularies, and disfluencies, can be addressed. In human user experimentation, finally, we will attempt to establish an assessment of whether simultaneous lecture translation is possible and if current output can lead to usable results. We will conclude that, while human interpreters are achieving a higher translation quality, automatic systems can already provide usable information for people who would otherwise be unable to understand the source-language lecture. Moreover, we will see that present system performance trails human performance by less than expected as human translators suffer from cognitive limitations in face of the real-time performance requirement.

1.1 Differences between interpreting and translating

Although the terms “translation” and “interpreting” are used interchangeably in everyday speech, they vary greatly in meaning. Both refer to the transfer of meaning between two languages. However, *translation* refers to the transfer of meaning from text to text, often with time and access to resources such as dictionaries, glossaries, etc. On the other hand *interpreting* consists of facilitating oral or sign language communication, either simultaneously or consecutively, between two or more speakers who are not speaking the same language.[1] The profession expects interpreters to be more than 80% accurate and translations, by contrast, over 99% accurate.[2]

Simultaneous interpreting is sometimes incorrectly referred to as “simultaneous translation” and the interpreter as the “translator”. However, in computer science the term “machine translation” (MT) is commonly used for systems translating text or speech from one language to another. The reason for this is that in the past the main focus of MT was the translation of text and only recently is SST attracting a wider interest. Therefore, throughout this paper we use the term *translation* as in “SST”, “simultaneous translation” or “simultaneous speech translation” to mean the automatic *interpretation* of spoken language.

1.2 Outline

The paper is structured as follows. In Sect. 2, challenges and advantages of human interpretation and automatic simultaneous translation are compared. Section 4 then explains a first baseline system for simultaneous translation, which is then further adapted to achieve a lower latency and better translation quality. The necessary techniques used for speaker and topic adaptation are explained in Sect. 5. Section 6 copes with the problem of finding a chunking of the speech recognizer’s output which achieves an optimal translation performance. Therefore, different chunking strategies are compared in an empirical study. The overall simultaneous translation system, its architecture and information flow is presented in Sect. 7, and Sect. 8 discusses the latency and real-time issues of such a system and possible solutions. Section 9 concentrates on the automatic and human end-to-end evaluation of the simultaneous lecture translator. Finally, Sect. 10 concludes the paper.

[1] Wikipedia entry “Interpreting”, http://en.wikipedia.org/wiki/Interpreting, accessed 20 June 2007.

[2] Wikipedia entry “Translation”, http://en.wikipedia.org/wiki/Translation, accessed 20 June 2007.

2 Human interpretation versus automatic simultaneous translation

Anyone speaking at least two different languages knows that translation and especially simultaneous interpreting are very challenging tasks. A human translator has to cope not only with the special nature of different languages like terminology and compound words, idioms, dialect terms or neologisms, unexplained acronyms or abbreviations and proper names, but also with stylistic differences and differences in the use of punctuation between two languages. And, translation or interpretation is not a word-by-word rendition of what was said or written in a source language; instead the meaning and intention of a given sentence has to be re-expressed in a natural and fluent way in another language.

2.1 Translating and interpreting in the EC

The majority of professional full-time conference interpreters work for international organizations like the United Nations, the European Union, or the African Union, whereas the world's largest employer of translators and interpreters is currently the EC with its two Directorate Generals for Translation and Interpretation.[3]

The Directorate General (DG) for Interpretation is the EC's interpreting service and conference organizer and provides interpreters for about 50–60 meetings per day in Brussels and elsewhere. The language arrangements for these meetings vary considerably, from consecutive interpreting between two languages, for which one interpreter is required, to simultaneous for 23 or more languages, which requires at least 69 interpreters. At present, the Council of the European Union accounts for around 46% of the interpreting services provided, followed by the Commission itself with around 40%. There are more than 500 staff interpreters, accompanied by 2,700 accredited freelance interpreters.[4]

In 2006, the European Parliament spent about €300 million, or 30% of its budget, on the interpretation and translation of the parliamentary speeches and EU documents. In total, approximately €1.1 billion are spent per year for the translating and interpreting services within the European Union, which is around 1% of the total EU budget (Accipio 2006, p. 23).

2.2 Challenges in human interpretation

According to Al-Khanji et al. (2000), researchers in the field of psychology, linguistics and interpretation like Henderson (1982), Hendricks (1971) and Seleskovitch (1978) seem to agree that simultaneous interpreting is a highly demanding cognitive task involving a basic psycholinguistic process. These processes require the interpreter to monitor, store and retrieve the input of the source language continuously in order to

[3] See http://europa.eu.int/comm/dgs/translation/index_en.htm and http://ec.europa.eu/scic, both accessed 29 October 2008.

[4] DG for Interpretation webpage "What we do", http://scic.ec.europa.eu/europa/jcms/c_5204/what-we-do-faq, accessed 20 June 2007.

produce the oral rendition of this input in the target language. It is clear that this type of difficult linguistic and cognitive operation will force even professional interpreters to resort to elaborate lexical or synthetic search strategies.

Fatigue and stress negatively affect the interpreter, leading to a decrease in interpreting quality. In a study by Moser-Mercer et al. (1998) in which professional speakers were told to work until they could no longer provide acceptable quality it was shown that (a) during the first 20 minutes the frequency of errors rose steadily, (b) the interpreters however appeared to be unaware of this decline in quality, (c) at 60 minutes, all subjects combined committed a total of 32.5 meaning errors, and (d) in the category of nonsense the number of errors almost doubled after 30 minutes on the task. In experiments with students and professional interpreters, Al-Khanji et al. (2000) found that the most frequent *compensatory strategies*, which are also used in such situations, are—in the order of occurrence—skipping, approximation, filtering, comprehension omission, and substitution. Therefore, in Vidal (1997), the conclusion was drawn that interpreters should work in teams of two or more and have to be exchanged every 30 minutes. Otherwise, the accuracy and completeness of simultaneous interpreters decreases precipitously, falling off by about 10% every five minutes after holding a satisfactory plateau for half an hour.

2.2.1 Fluency and the ear–voice span

Since the audience is able to evaluate the simultaneously interpreted discourse only by its form, the fluency of an interpretation is of utmost importance. According to a study by Kopczynski (1994, pp. 87–100), fluency and style was third on a list of priorities of elements rated by speakers and attendees that contribute to quality after content and terminology. Following the overview in Yagi (2000), an interpretation should be as natural and as authentic as possible, which means that artificial pauses in the middle of a sentence, hesitations, and false starts should be avoided (Jones 1998), and tempo and intensity of the speaker's voice should be imitated (Kopczynski 1994).

Another point to mention is the time span between a source-language chunk and its target-language chunk, which is often referred to as "ear–voice span", "delay", or "lag". Following the summary in Yagi (2000), the ear–voice span is variable in duration depending on some source- and target-language variables, like speech delivery rate, information density, redundancy, word order, syntactic characteristics, and so on. Nevertheless, the average ear–voice span for certain language combinations has been measured by many researches, but the range varies largely from two to six seconds (Barik 1969; Lederer 1978), depending on the speaking rate. Short delays are usually preferred for several reasons. For example, the audience is irritated when the delay is too large and is soon asking whether there is a problem with the interpretation. Another reason is that a short delay facilitates the indirect communication between the audience and the speaker but also between people listening to the interpreter and to the speaker. Therefore, interpreters tend to increase their speaking rate when the speaker has finished. In English, a latency of about two to six seconds is equivalent to a delay of about 4 to 12 words, since the average speaking rate is about 120 words per minute (Ramabhadran et al. 2003; Yuan et al. 2006).

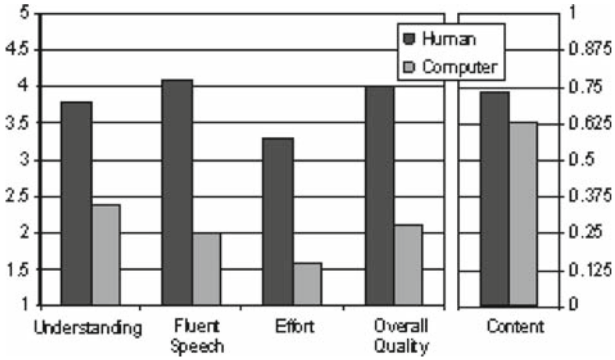

Fig. 1 Comparison of automatic and human interpretation performance judged by humans in the project TC-STAR

2.3 Advantages of a simultaneous translation system

Given the explanations above on human interpretation in general and in the EC in particular, one has to weigh two factors when considering the use of simultaneous translation systems: cost and translation quality. The comparative results of TC-STAR in Fig. 1 (Hamon et al. 2007) between human interpretation and automatic speech translation show that SST was judged worse when compared to human interpretation in most categories, but when it comes to the transfer of content both were judged nearly equal. The reason why even human interpretation is not reaching perfect results is due to the compensatory strategies mentioned above under which translation quality suffers. This means that for people not understanding the speaker's language at all such an automatic translation could be of great help. Furthermore, if there is additional information about, for example, the speaker or the topic of the speech available, adaptation techniques could be used to increase the overall performance of an automatic system. But in contrast to TC-STAR, where no real-time requirements were given, a simultaneous translation system has to deliver the translation with a latency as small as possible, and might therefore not achieve the same translation quality.

Another advantage of a simultaneous translation system compared to a human interpreter is the abundance of short-term memory: storing long sequences of words is not a problem for a computer system. Therefore compensatory strategies are not necessary, independently of the speaking rate of the speaker. However, depending on the system's translation speed it might not be able to keep up with the speaker, and, while it might be possible for humans to compress the length of an utterance without destroying the meaning (summarization), it is still a very challenging task for automatic systems (Mani 2001; Zechner 2002; Furui 2007).

At €300–400 per hour translator time, human simultaneous interpretation is quite expensive. Usually two interpreters are necessary and time for preparation and post-processing must be considered additionally. Furthermore, simultaneous interpretation requires a soundproof booth with audio equipment, which can be rented, but it adds overall cost that is unacceptable for all but the most elaborate multilingual events. On the other hand, a simultaneous translation system needs time and effort for preparation and adaptation towards the target application, language and domain. Depending on the required translation quality, the costs can therefore exceed those for a human interpretation. However, the major advantage of an automatic system is that once it is adapted it can be easily re-used in the same domain, language etc. A single laptop together with a microphone might already be sufficient.

To some extent, even generalization should not be a problem for automatic systems. Due to the manner in which such systems are trained, expressions not directly within the required domain, but closely related to it, are already covered by the system. Furthermore, adaptation techniques can be used to extend the coverage and quality, especially in situations where simultaneous translations into multiple languages are required, automatic systems are advantageous, because only the translation modules have to be extended with new target languages, while the rest of the system can be kept unchanged. Another advantage is that the transcript of a speech or lecture in the source and target languages is produced for free by using an automatic system. In the EU, for example, these transcripts can be used as an initial version of the protocols which have to be prepared anyway. This also allows for thinking about other possible delivering technologies beyond the traditional headphones.

Given the limitations of the recognition and translation capabilities of current SST systems and the system development costs compared to human interpreters, possible application scenarios for simultaneous speech translation are restricted to domains to which a system can be well adapted and applications in which a system can be re-used and modified or customized with less effort. Therefore, we selected the lecture scenario as our target scenario, in which a single speaker is talking about a specific topic to a larger audience. Hence, small talks, student seminars or parliamentary speeches also belong to this scenario. However, it should be noted that finding an optimal balance between quality and speed is always a difficult task to undertake.

3 Spoken language translation

A spoken language translation system always consists of the two main components, an automatic speech recognition (ASR) system and an MT system. In addition, other components are typically required as for example for preparing the speech recognizer's output for the MT system in order to achieve high translation quality. Output may or may not be in the form of speech, in which case speech synthesis is also invovled.

Figure 2 gives a schematic overview of the simultaneous translation architecture treated in this paper together with required databases and models. From the continuous input stream of speech, the ASR component produces a continuous stream of partial first-best hypotheses, which is resegmented into appropriate chunks for the MT component. The MT component translates each of these source-language chunks into the target language. By using multiple MT components translation can be done in parallel into different target languages at the same time. For delivering the translation output, different technologies may be used among which the most prominent are either subtitles or speech synthesis. A more detailed description will be given in Sect. 7.

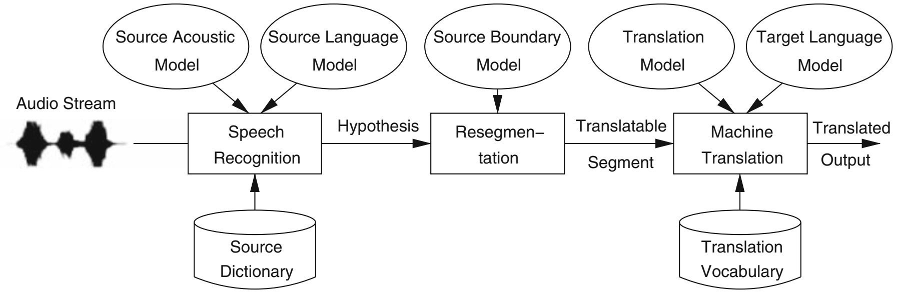

Fig. 2 Schematic overview of the simultaneous translation architecture treated in this paper. Boxes re-present the components of the system, and ellipses the models used by the components. As databases, dictionaries and vocabularies are used

3.1 Speech translation projects

In the following, some related research projects such as STR-DUST, TC-STAR, and GALE are described in more detail.

STR-DUST (Speech TRanslation: Domain-Unlimited, Spontaneous and Trainable),[5] an NSF-ITR project launched in 2003 at Carnegie Mellon University, aimed to address the problem of domain independence. The goal of the project was to explore translation under the added challenge of domain independence and spontaneous language. The project studied summarization and disfluency removal techniques for disfluent spoken language, such as spontaneous lectures and meetings, and integrated them in domain-unlimited spontaneous translation systems. It focused on highly disfluent speech such as dialogs and spontaneous speech without domain restrictions. It also explored improving and training open-domain speech translators using human interpreter speech.

TC-STAR (Technologies and Corpora for Speech-to-Speech-Translation),[6] an EC-financed project that started in April 2004 within the "6th Framework Programme", was a large-scale, long-term effort to advance the performance of all necessary core technologies for high-performance domain-independent SST, i.e. ASR, MT and text-to-speech (TTS). The objective of the project was to make a breakthrough in SST that significantly reduces the gap between human and machine translation performance. The focus was on the development of new algorithms and methods. The project achieved these targets in only three years, by reducing the ASR error rate from 30% to 7%, and establishing understandable end-to-end translation output. It was a communal project of several partners, and data recordings, transcriptions or speeches from the European Parliament, and broadcast news were used. Systems in three languages were developed: European English, European Spanish, and Mandarin Chinese.

The goal of the DARPA GALE (Global Autonomous Language Exploitation) program[7] is to develop and apply computer software technologies to absorb, analyze and

[5] See http://isl.ira.uka.de/index.php?id=17, accessed 29 October 2008.

[6] See http://www.tc-star.org, accessed 29 October 2008.

[7] See http://www.darpa.mil/ipto/programs/gale/gale.asp, accessed 29 October 2008.

interpret huge volumes of speech and text in multiple languages. Automatic processing engines will convert and distill the data, delivering pertinent, consolidated information in easy-to-understand forms to military personnel and monolingual English-speaking analysts in response to direct or implicit requests. In contrast to TC-STAR, the output of each engine is English translated text only and no speech synthesis is used. Instead, a distillation engine is responsible for integrating information of interest to its user from multiple sources and documents. The input to the transcription engine is speech, currently with a main focus on Arabic and Chinese. Military personnel will interact with the distillation engine via interfaces that could include various forms of human–machine dialog (not necessarily in natural language).

The focus in these projects was to achieve high-quality translations without paying attention to the latency or processing time. As a result, the architectures of the systems developed within these projects are more complicated compared to the architecture presented above. Within TC-STAR, individual system components developed by the project partners were combined at both the ASR and MT stage into a distributed TC-STAR system. This resulted in significant performance improvements with respect to performance of the single systems. For ASR, several transcription systems were combined based on ROVER (Fiscus 1997). The final TC-STAR system output was then generated by combining the output of up to six state-of-the-art statistical phrase-based translation systems from different project partners into a consensus translation. The individual systems used different additional features in log-linear model combination, including continuous space language models, local reordering based on manually created or automatically learned part-of-speech templates, nonlocal reordering based on parsing, word disambiguation using morphosyntactic information, clustered language models and sentence mixture language models that model topic dependency of sentences, bilingual *n*-gram translation models, and additional techniques such as punctuation prediction, *n*-best list rescoring, and morphology reduction in preprocessing. More details on the individual systems as well as on the system combination algorithm can be found in Matusov et al. (2008).

In GALE, similar techniques are used for system combination.

3.2 Performance measures

In the following, some automatically computable performance measures will be described briefly. They are used to judge the quality and speed of the subcomponents of the simultaneous translation system, as well as the system's overall latency. Compared to human judgement, automatically computable performance measures are preferred especially during system development, to accelerate the turn-around time of the necessary experiments.

Perplexity: The perplexity is a measure of language model performance and is based on the entropy, a measure of uncertainty associated with a random variable in the field of information theory.

OOV rate: The OOV (out-of-vocabulary) rate is defined as the percentage of word occurrences in a corpus which are not covered by the vocabulary.

WER: The word error rate (WER) is a standard metric to measure the quality of an ASR system. It is derived from the Levenshtein or minimum edit distance and is defined as the minimum edit distance between a given hypothesis and its reference transcription normalized by the length of the reference transcription in words.

Bleu: The Bleu score (Papineni et al. 2002) together with the NIST score (see below) are standard metrics for automatically measuring the quality of an MT system. Both are based on the idea of a modified n-gram precision based on n-gram co-occurrence statistics. Bleu is defined as the geometric mean of modified n-gram precision scores p_n multiplied by a brevity penalty.

NIST: Compared to the Bleu score, the NIST score (Doddington 2002) mainly differs in two characteristics. First, it uses an information criterion via which n-grams that occur less frequently, and are therefore more informative, are weighted higher. Second, it differs in the brevity penalty factor insofar as the impact of small variations in the translation length is minimized.

RTF: The real-time factor (RTF) describes the ratio between the duration d of an input and the time p necessary to process that input. The RTF is always machine dependent and is always computed in this work on an Intel Pentium D with 3GHz and 4GB of memory running under SuSE Linux 10.0. The Janus executable was compiled with the Intel C++ Compiler v9.1 using auto-vectorization

Latency: The latency describes the delay of a system between the input at time i and the processed output of the given input at time o. In this work, the latency describes the delay until a given speech segment is recognized, translated, and output, and will be measured in words or seconds.

Precision, Recall, and f-measure: Precision and recall are two measures widely used to evaluate the quality of information retrieval systems. In this context, recall describes the completeness of a search result and precision its accuracy. In the interest of having a single performance measure, the f-measure is defined as the weighted harmonic mean of precision and recall.

It is known from the literature (Klakow and Peters 2002) that there is a positive almost linear correlation between perplexity and WER. The WER also depends on the OOV rate. As a rule of thumb, it can be said that an occurrence of an OOV word is on average responsible for 1.5 to 2 errors (Geutner et al. 1998).

From results obtained within TC-STAR and from our studies in Stüker et al. (2007) it can be seen that there is an negative almost linear correlation between the WER and MT quality.

Although it was shown in Snover et al. (2006) that automatic measures for MT quality correlate with human judgment, current automatic measures are only to some extent able to judge semantic differences. This means that errors produced by MT as well as ASR which affect a few words only, but destroy the semantic meaning of the whole sentence, are underestimated by all automatically computable performance measures. Therefore, the end-to-end performance of the simultaneous translation system will be also judged by humans in Sect. 9.

Table 1 Summary of the data used for acoustic model training

Type	Meetings		Speeches	Lectures	
	ICSI	NIST	EPPS	TED	SMNR
Duration (hrs)	72	13	80	10	10
Speakers	53	61	1894	52	67
Recordings	75	19	63	39	17

4 A first unadapted baseline system

This section introduces a first unadapted baseline system for SST of lectures and speeches.

Since lectures and speeches are usually held in rooms with a large audience, microphones used at a small distance from the speaker's mouth, like head-worn or directional stand microphones are preferred over far-field microphones which are more sensitive to background noise. Therefore, the speech recognition systems were developed and optimized with respect to close-talking recording conditions. Since this research evolved in the context of the European projects CHIL and TC-STAR, European accented English was of most interest.

4.1 Training, test, and evaluation data

This section describes the data used for training the acoustic, translation and language models for the ASR and MT systems. Furthermore, it presents the details of the development data as well as the evaluation data used to tune the systems and for performing the following experiments.

4.1.1 Acoustic model training data

As already mentioned, spontaneous close-talking European-accented English is of most interest. Therefore, for acoustic model training, recordings from meetings, European Parliament speeches, lectures and student seminars were selected.

The most suitable data are the lectures and seminars collected in the CHIL project (Waibel et al. 2004) (SMNR) and the Translanguage English Database (TED) (Lamel et al. 1994). Since the amount of data is insufficient to train acoustic models that perform well in large vocabulary speech recognition tasks such as lecture recognition, other data has to be used as well. The meeting data consists of recordings of highly spontaneous meeting speech from mostly native English speakers collected at two different sites: the International Computer Science Institute (ICSI), Berkeley, and the US Government's National Institute of Standards and Technology (NIST). In contrast, the manually transcribed European Parliamentary plenary sessions (EPPS) covers nonnative speeches, which are usually prepared in advance and therefore less spontaneous, i.e. planned speech. An overview of the details of the several training corpora is given in Table 1.

Table 2 Summary of the bilingual training data used for MT

	English	Spanish
Sentences	1,162,176	
Words	27.7M	28.9M
Vocabulary	93,157	130,473
Singletons	34,543	45,400

4.1.2 Translation model training data

The unadapted translation system was trained on a corpus of sentence-aligned parallel final text editions from the European Parliament available through its website. The data was crawled, preprocessed and sentence-aligned automatically by RWTH Aachen in TC-STAR (Gollan et al. 2005). The corpus statistics of the preprocessed training corpora are shown in Table 2.

4.1.3 Language model training data

The language models for the baseline speech recognition systems were trained on the corpora used for the NIST 2006 Rich Transcription evaluation (RT-06S) (Fügen et al. 2006a) and the 2007 TC-STAR evaluation systems (Stüker et al. 2006). Three different types of data were used: speech transcripts, written text, and text data collected from the World Wide Web. An overview of the data is given in Table 3. For training the language models of the MT systems the same bilingual data was used as described above in Sect. 4.1.2.

4.1.4 Development and evaluation data

For development and evaluation, a total of 19 lectures and talks on different topics by a single nonnative speaker were collected in different environmental conditions. All lectures are university lectures related to the field of speech and language processing. The talks were given for different reasons such as keynotes, press conferences, project reviews, or just an overview about the work done at our research lab. Table 4 gives an overview of the data. While lectDev and lectEval were used for system development and evaluation, lectOther was used for the adaptation experiments. For the development of the MT system, two talks out of lectDev were translated into Spanish, namely t035 and t041, with a duration of 44 minutes and 7,498 words in English. For the evaluation of the MT and for measuring the end-to-end performance of the simultaneous translation system, only a subset of lectEval was translated from English into Spanish. It consists of the two talks t036+ and l043 totalling 92 minutes and 14,908 words.

In addition, some speech recognition experiments were performed on the official development and evaluation sets of the TC-STAR-07 EPPS evaluation (EPPSdev, EPPSeval), further referred to as “parliamentary speeches”, and the lecture meeting data track of the NIST RT-06S evaluation (RT-06Sdev, RT-06Seval), further referred to as “lecture meetings”. Especially the lecture meetings are well suitable as an additional development set because of their similarity in speaking style and focus

Table 3 Overview of the available English language model training data

Type	ID	Words	Description and source
Parliamentary	EPPS-S	750k	Manual transcriptions (Gollan et al. 2005)
Speeches	EPPS-T	33M	Final text editions (Gollan et al. 2005)
	EPPS-U	1.4M	Automatic transcriptions (Gollan et al. 2007)
Meetings	MTG	1.1M	Meeting transcriptions (Burger et al. 2004; Garofolo et al. 2004; Janin et al. 2004)
	AMI	200k	Meeting transcriptions from Augmented Multi-party Interaction (AMI) (Carletta et al. 2005)
Talks, seminars	TED	98k	Translanguage English Database (Lamel et al. 1994)
	SMNR	45k	Transcriptions from CHIL (Waibel et al. 2004)
Conversations	SWB	4M	Transcriptions of from Switchboard (Godfrey and Holliman 1993)
News texts	BN	131M	Broadcast news texts (Graff et al. 1997; Fiscus et al. 1998)
	UN	42M	UN parallel corpus v1.0 (Graff 1994)
	HNSRD	48M	UK Parliament debates (Roukos et al. 1995)
	GWRD	167M	Extracts from the Gigaword corpus (Graff 2003)
Proceedings	PROC	23M	Recent related conference proceedings
Web data	UW-M	147M	From UWash, related to MTG
	UKA-M	124M	Related to MTG and PROC
	UKA-L	146M	Related to SMNR
(Query based	UKA-LP	130M	Related to SMNR and PROC
filtered)	UKA-MP	124M	Related to MTG and PROC

Table 4 Development and evaluation data

Set	Talks	Duration (min)	Words
lectDev	6	243	40,405
lectEval	5	172	24,104
lectOther	8	295	47,751

on a particular topic. However, the level of detail is even higher than the average in lectDev and lectEval, and many of the recorded speakers have a strong foreign accent. Compared to the talks and lectures introduced above, both characteristics are responsible for a more difficult recognition.

4.2 Speech recognition

Interest in research in the domain of the automatic transcription of lectures and speeches has grown, especially in the last couple of years. To date, several research groups and projects across the world deal with this topic. Within the EC-financed project, CHIL, international evaluations on lecture meetings were carried out in the context of the 2006 NIST Rich Transcription evaluation (Lamel et al. 2006; Huang et al. 2006; Fügen et al. 2006a). Besides this, other research has been conducted, for example on the TED corpus (Lamel et al. 1994; Cettolo et al. 2004). The focus of the European funded project TC-STAR was the transcription of European Parliament plenary Sessions (Ramabhadran et al. 2006; Lööf et al. 2006; Stüker et al. 2006; Lamel et al. 2007). LECTRA (Trancoso et al. 2006) was a national Portuguese project focusing on the production of multimedia lecture contents for e-learning applications, which implies taking the recorded audiovisual signal and adding the automatically produced speech transcription as caption. The goal of the MIT Spoken Lecture Processing project (Glass et al. 2007) is to improve access to on-line audio/visual recordings of academic lectures by developing tools for the processing, transcription, indexing, segmentation, summarization, retrieval and browsing of this medium. In the Liberated Learning Consortium (Bain et al. 2005), instead the goal is to provide real-time, high-quality automatic speech transcription to aid hearing-impaired students. Research is conducted not only on improving the recognition quality, but also on the delivery aspect. Real-time processing is also required in the FAME project (Rogina and Schaaf 2002), where a lecture and presentation tracker is presented. The biggest research effort on a single spoken lecture corpus has been in Japan using the Corpus of Spontaneous Japanese, which consists of about 700 hours of speech (Furui 2005).

In this section, a first unadapted baseline speech recognition system for lectures and speeches is introduced. As a starting point the evaluation systems built for RT-06S and TC-STAR-07 were used. Typically, the evaluation conditions do not limit the time for processing the audio data, whereby the speech recognition systems were designed to achieve an optimal recognition accuracy rather than a short decoding time. This was also the case in the above-mentioned evaluations, RT-06S and TC-STAR-07, where the final ASR result was computed by combining the results of different multipass decoding and adaptation chains using different kinds of acoustic and language models. In contrast to this, simultaneous translation requires that all components run in real time with low latency. Thus, a single-pass decoding system which meets these requirements had to be designed.

From an acoustic model point of view, optimization was done with respect to the WER on the lecture meeting data track of RT-06S and the parliamentary speeches of the 2007 TC-STAR evaluation (TC-STAR-07) mentioned in Sect. 4.1.4, in addition to the lecture development data used in this paper. This was done since lectures and speeches vary greatly in their spontaneity and we would like to cover both highly spontaneous speech as in the lecture meeting data track and less spontaneous speech as in the parliamentary speeches. In deference to this goal, the language models were tuned with respect to the specific domain. Furthermore, this development strategy allows for a direct comparison with other existing evaluation systems within the research community.

Acoustic model training was done with the help of the Janus Recognition Toolkit (JRTk) (Finke et al. 1997), which was jointly developed at Carnegie Mellon University, Pittsburgh, PA and Universität Karlsruhe (TH), Germany. For decoding and lattice generation, the single-pass Ibis decoder (Soltau et al. 2001), developed at Universität Karlsruhe (TH) and also part of the JRTk, was used. The language models used for

Table 5 In- and across-domain perplexity and OOV rate comparison on the different development sets

Development set	RT-06S		TC-STAR-07	
	OOV	Perplexity	OOV	Perplexity
RT-06Sdev	1.09	148	2.42	283
EPPSdev	1.35	250	0.60	84
lectDev	0.47	160	1.94	255

decoding and lattice rescoring were built using the SRI Language Modeling Toolkit (Stolcke 2002).

4.2.1 Vocabulary, dictionary, and language model

As expected, the results in Table 5 show that the vocabulary and language model from the RT-06S evaluation system are more suitable for lecture recognition than the components from the TC-STAR-07 system. Therefore, the following short description concentrates on the components of the RT-06S system. More details can be found in Fügen et al. (2006a).

The dictionary contained 58.7k pronunciation variants over a vocabulary of 51.7k words. Pronunciations were either derived from already existing dictionaries or automatically generated using Festival (Black and Taylor 1997).[8] The language model used in the RT-06S system is a 4-gram mixture language model, whereas the mixture weights for each language model were optimized on a held-out set with a size of 30k words. The language model components were trained on the following corpora: meeting data transcripts (MTG, AMI), transcripts of lectures and seminars (SMNR, TED), text data (BN, PROC), and several data sets collected from the Web (UW-M, UKA-L, UKA-LP, UKA-MP). Besides the standard framework for Web text collection proposed in Bulyko et al. (2003), we followed also another approach, where the frequent n-grams from different lecture or conference meeting transcripts were combined with topic bigrams from the conference proceedings of PROC to form queries. The topic phrase generation consisted of: computing bigram tf-idf (term frequency–inverse document frequency) weights for each document in the proceedings data, zeroing all but the top 10%, averaging these weight vectors over the collection, and taking the top 1,400 bigrams excluding any with stop-words or numbers. The topic bigrams were mixed randomly with the general phrases until the desired number of queries (14k) were generated.

4.2.2 Acoustic model

Based on the experimental results in Fügen et al. (2006b), it was seen that relatively good results on all development sets could be achieved when using SMNR, ICSI, NIST, TED, and EPPS-S data for acoustic model training. The training procedure

[8] See also http://www.cstr.ed.ac.uk/projects/festival, accessed 29 October 2008.

described in Soltau et al. (2004) was extended by applying discriminative training (Bahl et al. 1986; Normandin 1991) and speaker adaptive training using constrained maximum likelihood linear regression (MLLR) (Gales 1998). Altogether, the first unadapted system was trained on nearly 190 hours of speech data, resulting in a fully continuous acoustic model with approx. 234k Gaussians divided amongst 4,000 context-dependent models in a 42-dimensional feature space. For feature extraction, the same method as described in Fügen et al. (2006a) was used. A fully continuous acoustic model was preferred over a semi-continuous model, because of the better ratio between speed and accuracy.

4.2.3 Results of the first unadapted system

For the first unadapted baseline system, the multipass decoding strategy used for RT-06S and TC-STAR-07 was reduced to the first decoding pass only. The final result can be seen in Fig. 3. For RT-06S and TC-STAR-07, the matching vocabulary and language model was used, and for the lectures, the components from RT-06S. In all cases, incremental online adaptation per speaker or lecture as described in Fügen et al. (2006a) and Stüker et al. (2006) was used. Furthermore, the impact of language model rescoring on the WER is analyzed, and the mean WER and its standard deviation per speaker in a lecture meeting, parliamentary speech, or lecture is shown. It can be seen that while parliamentary speeches and lectures can be recognized with an almost equal recognition accuracy and speed, the lecture meetings are much more difficult. Since the lectures were collected from a single speaker only, the variance in terms of WERs is very low compared to the variance for the parliamentary speeches and lecture meetings, but especially for TC-STAR-07, a significant number of speakers with WERs lower than those obtained on the lectures exist. It should be noted that for the lectures the mean WER and its standard deviation was computed across the different lectures of the same speaker, but for RT-06S and TC-STAR-07, across different speakers and lecture meetings or parliamentary speeches.

In anticipation of Sect. 8, it should be mentioned that although the Ibis decoder is a single-pass decoder making extensive use of language model information with arbitrary context length, already during decoding it can be seen that lattice-based language model rescoring consistently lowers the WER. It is advantageous that the RTF is only marginally influenced by the additional time spent for lattice generation and rescoring.

4.3 Machine translation

In this section, we discuss the approach for developing the statistical MT (SMT) component in our lecture translator that was used to translate lectures and speeches from English to Spanish. The initial purpose of the underlying phrase-based SMT system developed within TC-STAR was to translate speeches from EPPS in an offline scenario, that is, translating without constraints with respect to some parameters which are critical in the simultaneous speech translation scenario, such as processing time, resource usage, and latency. Parliamentary speeches, while covering a relatively broad

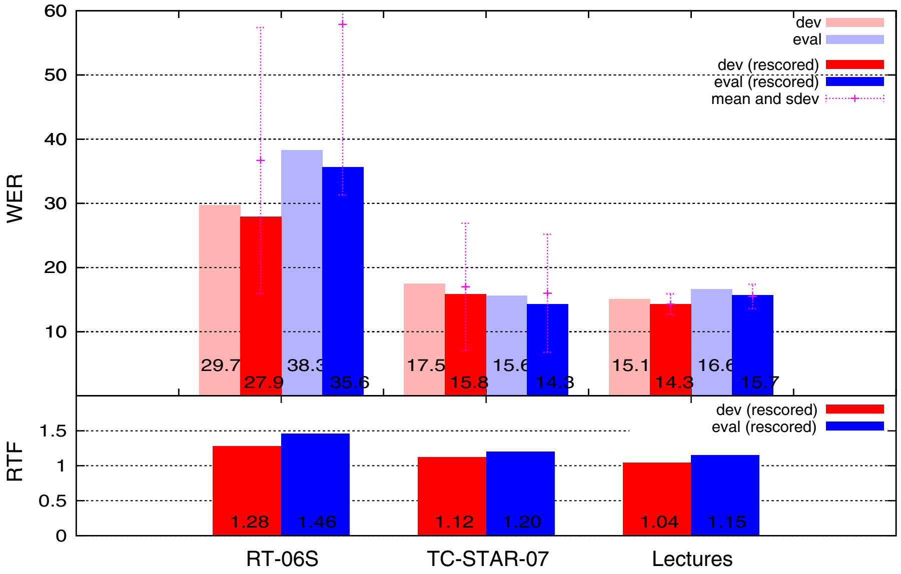

Fig. 3 Comparison of WERs on the different development and evaluation sets before and after language model rescoring in combination with the RTFs computed with language model rescoring. Furthermore, the lines within the WER boxes mark the mean WER and its standard deviation per speaker in a lecture meeting, parliamentary speech, or lecture

discourse domain, are usually prepared speeches given by well-trained speakers. Lectures, on the other hand, can go into much more detail on any given topic, and, as an aggregate, cover a practically unlimited domain; a system suitable for translating general lectures must be able to cope with much more variable and more spontaneous speaking styles. At the same time, the MT component should make use of the additional information about the topic, which is known in advance in the typical lecture translation scenario, and offer acceptable speed and latency in real-time operation. We therefore needed to develop a system considerably different from the one built and optimized for traditional batch mode operation and well-controlled speech. In our experiments, we used loose coupling, passing the single-best hypothesis from the recognizer to the translation component. Translation quality was measured using the well-known evaluation metrics Bleu (Papineni et al. 2002) and NIST (Doddington 2002). MT scores were calculated using case-insensitive scoring and one reference translation per test set.

4.3.1 Word and phrase alignment

Phrase-to-phrase translations are critical for the performance of state-of-the-art SMT systems, and our lecture translation system builds upon this foundation. For the initial version of the MT component, we chose an approach to phrase alignment and extraction which is well suited for efficiently providing phrase translation alternatives on the fly (Vogel 2005). This method, based on optimizing a constrained word-to-word alignment for an entire sentence pair, can be used to extract phrase translation candidates of arbitrary length from the training corpus at decoding time (Kolss et al. 2006). To find a translation for a source phrase, the precomputed word alignment is restricted: words inside the source phrase align to words inside the target phrase, and words outside the source phrase align outside the target phrase. This constrained alignment probability is calculated using the well-known IBM Model-1 word alignment model (Brown et al. 1994), but the summation of the target words is restricted to the appropriate regions in the target sentence. The position alignment probabilities are adjusted accordingly.

For the experiments reported in this paper, we used a classic phrase table constructed by training IBM Model-4 word alignments in both directions and extracting phrase-to-phrase translation pairs which are consistent with these word alignments, as this significantly improved translation quality. The GIZA++ toolkit (Och and Ney 2003) and the implementation of the "grow-(diag)-final" heuristic (Koehn et al. 2005) were used for training word and phrase alignments, respectively. In addition, we applied modified Kneser-Ney discounting to the raw relative frequency estimates of the phrases as described by (Foster et al. 2006). Finally, the phrase table was pruned to the top ten phrase translations for each source phrase using the combined translation model score as determined by Minimum Error Rate (Och 2003) optimization on a development set.

4.3.2 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 as described in the previous section.
2. A 4-gram language model trained with the SRI language model toolkit (Stolcke 2002).
3. A word reordering model, which assigns higher costs to longer-distance reordering. We use the jump probabilities $p(j|j', J)$ 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 lying or phrases starting within

Table 6 MT results on text as well as on ASR input

	NIST	Bleu
t036+, text input	5.72	23.4
t036+, ASR input	5.06	17.9
l043, text input	5.27	19.6
l043, ASR input	4.80	16.6

a given look-ahead window from the current position (Vogel 2003). The usage of a local reordering window in the lecture translation system captures most of the benefits of improved target word order while keeping search complexity low for real-time operation.

4.3.3 A first unadapted system

For training the baseline translation system, the parallel EPPS corpus as described in Sect. 4.1.2 was used. Although the speech recognition is unable to deliver punctuation marks with its hypotheses, they were left in the corpus but separated from the words. The reason for this is that they help in finding useful split points for the phrase alignment training. Misalignments in the corpus and other "noise" like document references were discarded. Furthermore, abbreviations were expanded and dates and number expressions spelled out. In conformity with the speech recognition all data was lower cased.

All punctuation marks were later removed from the source side of the phrase alignments to keep the interface between ASR and MT as simple as possible. As later explained in Sect. 6 developing an additional punctuation annotation module with a reasonable performance is quite challenging, and according to Matusov et al. (2006) the performance of the SMT is affected only slightly.

For training the initial language model, the target, i.e. Spanish, side of the same corpus was used. Modified Kneser-Ney discounting and interpolation of discounted n-gram probability estimates was applied. This results in a trigram language model covering 131k words and having a perplexity on the Spanish side of `lectDev` of 631.

Table 6 summarizes the translation results of the first baseline system on the development and evaluation data for the lectures and TC-STAR-07. As expected, the performance on ASR results is worse compared to that on manually transcripts. It should be noted that from results obtained within TC-STAR and from studies done by ourselves in Stüker et al. (2007) it can be seen that there is an negative almost linear correlation between the WER and the MT quality.

5 Adaptation

It is a known fact that adaptation increases a system's performance, and lectures and speeches are particularly suited for adaptation because (a) a speaker is talking more or less continuously for a longer time, (b) the identity of the speaker who gives the lecture is known in advance, and (c) the topic of the lecture is known in advance as well. This makes it possible to perform adaptation on different levels or with respect to different models of a simultaneous translation system:

Acoustic model: Acoustic model adaptation can be used to adapt to a speaker's voice or to compensate for variations in the audio signal, like ambient noise.
Source-language model: Language model adaptation can be used to adapt a language model to a specific talk or lecture, but also to the speaker's speaking style, which can differ in the degree of spontaneity as well as in the selection of phrasing.
Translation model: Translation model adaptation can improve the translation quality in all situations where the training data does not match the new conditions and is therefore used mainly for topic adaptation especially when new, unseen phrases or words have to be translated.
Target-language model: The target-language model is used by the MT system and can be adapted for the same reasons and in the same way as the source-language model.

In general, adaptation techniques fall into two categories: normalization of the input to match the model to which the system was trained to, and model adaptation techniques in which the parameters of the model itself are adjusted to match the input better. A further distinction can be made between *supervised adaptation*, where the adaptation parameters are estimated with respect to the reference transcriptions, and *unsupervised adaptation*, where hypotheses are used instead. Some of the techniques can be used during acoustic model training only (*offline adaptation*), others also during decoding for incremental adaptation (*online adaptation*).

Adaptation is not special to automatic systems only; human interpreters do that as well. Especially for simultaneous interpretation human interpreters interview the speaker and carry out some literature research in advance in order to make themselves familiar with the topic. In this way, the translation of special phrases can be memorized in order to allow for a more fluent interpretation.

5.1 Speaker and topic adaptation for ASR

5.1.1 Speaker and channel adaptation

Because of their simplicity, cepstral mean and variance normalization (Atal 1974; Furui 1986), vocal tract length normalization (VTLN) (Eide and Gish 1996; Zhan and Westphal 1997) and constrained MLLR (cMLLR) (Gales 1998) are particularly suited for online speaker and channel adaptation. Compared to no adaptation at all, achieving a WER of 16.1% on `lectDev` and 19.8% on `lectEval` using incremental VTLN and cMLLR in addition decreases the WER to 14.1% on `lectDev` and even 15.5% on `lectEval`.

An insight into the behavior and improvements of online adaptation, given in Fig. 4, shows the WERs obtained for some selected talks that are plotted over the number of utterances decoded so far. It can be seen that the WER is continuously going down over time.

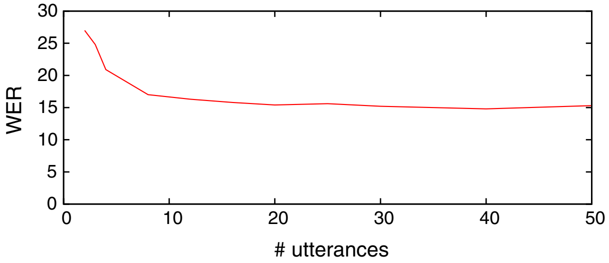

Fig. 4 Improvements of incremental adaptation with respect to the number of utterances available for parameter estimation

Figure 5, in contrast, shows the results for supervised and unsupervised MLLR offline adaptation (Leggetter and Woodland 1995). It can be seen that the WERs are decreasing with an increasing amount of adaptation data, whereas the difference in WER between supervised and unsupervised adaptation is getting larger with increasing availability of adaptation data. Furthermore, the additional gain due to incremental online adaptation using the precomputed offline adaptation parameters as a starting point can be seen as well. At a reasonable amount of adaptation data of 15 minutes which usually corresponds to an average talking time of a single talk, a relative improvement in WER on `lectDev` of almost 5.7% after supervised and 2.1% after unsupervised adaptation could be achieved. When using the full set, the relative improvement increases to 12.8% after supervised and 10% after unsupervised adaptation. In another experiment, it was determined whether the difference in WER between supervised and unsupervised adaptation can be preserved when the adaptation parameters and acoustic models are seeded with that of the supervised adapted system. Starting from a supervised adapted system using 15 minutes of speech data, unsupervised adaptation was performed. Unfortunately, the difference in WER could not be preserved, as can be seen from the "mixed" curve in Fig. 5. Another interesting observation from the experiments is that when using no online adaptation during decoding, a system adapted with a small amount of adaptation data actually performs worse as compared to an unadapted system using online adaptation. In the other case, i.e. with online adaptation during decoding, even a small amount of adaptation data leads to an improvement in WER of the adapted system.

5.1.2 Topic adaptation

In the RT-06S evaluation system, topic adaptation was already performed by collecting Web data related to scientific lectures as well as using proceeding data in addition. A mixture model approach was used to incorporate that data into the final language model by training single language model components on each of the training corpora and interpolating them together into a single language model. Hence, the interpolation weights for the different language model components are computed iteratively using the EM algorithm by minimizing the perplexity on a held-out set (de Mori and Federico 1999; Bellegarda 2004).

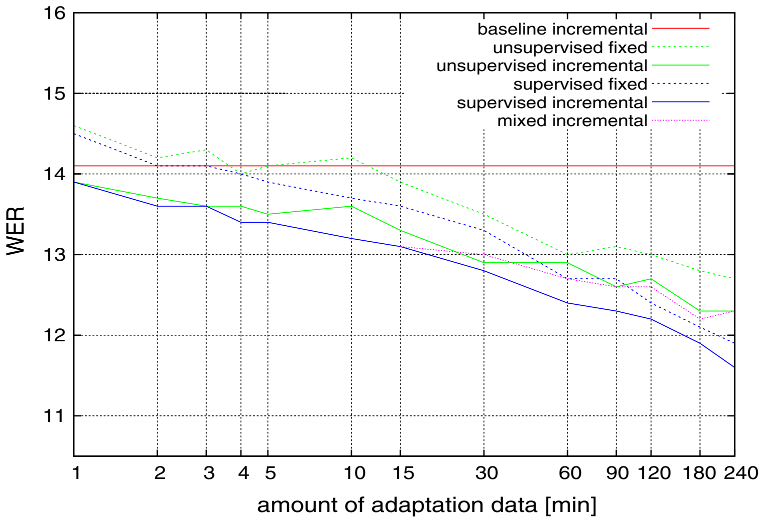

Fig. 5 Comparison of supervised and unsupervised MLLR offline adaptation with respect to the amount of available data. For the curves marked "incremental", online adaptation was performed during decoding, while for the curves marked "fixed" the precomputed offline adaptation parameters were used instead and no online adaptation was performed. For the "mixed" curve the unsupervised adaptation was seeded with a supervised adapted system

Figure 6 shows the interpolation weights for different language models and gives an impression of the relevance of the different corpora for specific domains. The columns `RT-06S` and `TC-STAR-07` show the weights for the language models used in the respective evaluations. In contrast to `TC-STAR-07`, where the weights were optimized on the development set itself (`EPPSdev06`), the weights for `RT-06S` were optimized on a separate held-out set. Therefore, the next column `RT-06Sdev` shows the interpolation weights optimized on `RT-06Sdev` instead. The most obvious change is the increased weight of the meeting component (`MTG`) together with the decreased weight of the seminar component (`SMNR`). This clearly shows the difference between the held-out set containing lectures only and the development set containing more interactive lecture meetings.

In the last three columns marked with `-all`, the interpolation weights were optimized using all available data and not restricted to the subset available at the time of evaluation or allowed by the evaluation conditions. It should be noted that for reasons of clarity only the interpolation weights of the language model components which contribute more than 3% to the mixed language model are shown. It can be seen that for `RT-06S`, in addition to the components used so far, the `EPPS-S` and `UKA-L` components were used. For `TC-STAR-07`, only the `UKA-L` Web data component was added. Due to the lack of matching in domain data, it can be seen that for lectures, (`lectDev-all`) more corpora are used than compared to the other domains and that the most relevant corpora were `UKA-MP`, `PROC`, `EPPS-S`, and `MTG`. When adding transcribed in-domain lecture data (`lectOther`) from the same speaker this completely changes as can be seen in the last column `lectDev-lect`. The weight

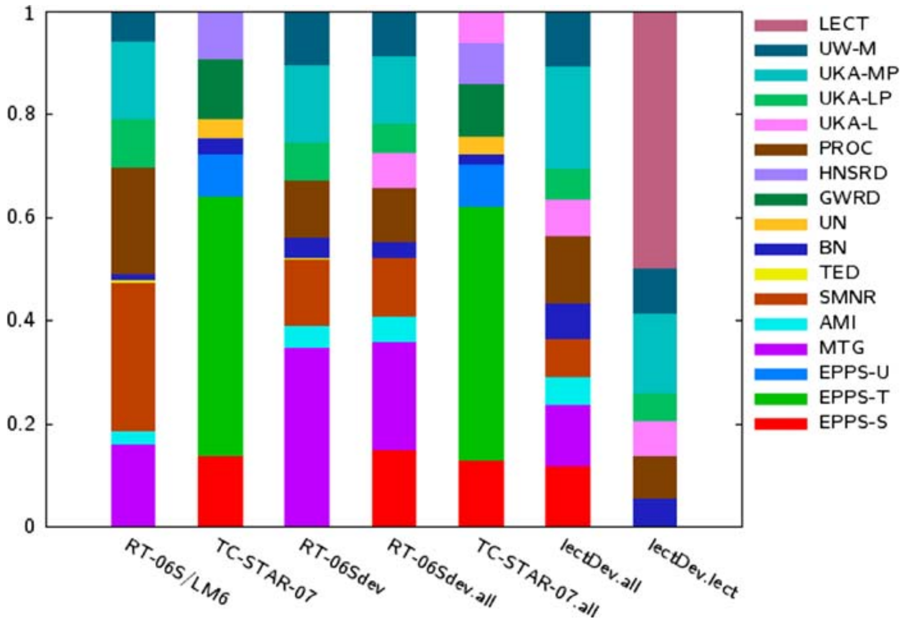

Fig. 6 Interpolation weights for different mixture language models optimized on different development sets. For each language model component, the corpus used for training is given in the legend and was described in Sect. 4.1.3. `RT-06S` and `TC-STAR-07` are the weights of the original language models used in the evaluation system; for the other columns the weights were recomputed using the additional corpora as well

Table 7 In- and across-domain perplexity comparison of language models tuned on different development sets

Language model	RT-06Sdev	EPPSdev	lectDev
RT-06S	153	276	163
TC-STAR-07	283	84	255
RT-06Sdev	148	250	160
RT-06Sdev.all	134	150	148
EPPSdev.all	227	81	209
lectDev.all	140	147	151
lectDev.lect	136	208	117

of the lecture data is almost 50% and shows the clear match in topic and speaking style.

In Table 7, the perplexities for the different language models on the different sets are summarized. It can be seen that due to tuning of the language model weights, the perplexity can be decreased from 163 to 151 on `lectDev` and by adding in-domain lecture data further down to 117.

Table 8 Comparison of the OOV rates (%) of the original RT-06S vocabulary used in the baseline system and the expanded vocabulary on the lecture data

	RT-06S	Expanded
lectDev	0.47	0.14
lectOther	0.46	0.15
lectEval	0.74	0.49

Table 9 Comparison of the final recognition results (percentage WERs) after language model rescoring using the techniques described above on lectDev and lectEval

	lectDev	lectEval
No adaptation	16.4	19.9
Acoustic model online adaptation	14.3	15.7
+ Offline adaptation	12.7	14.0
+ Language model adaptation	12.0	13.4
+ Vocabulary expansion	11.2	12.6

5.1.3 Vocabulary expansion

In Table 8, it can be observed that the OOV rate with the RT-06S vocabulary used in the baseline system is pretty low already. Nevertheless, it is still worth decreasing the OOV rate by expanding the vocabulary even more, as will be seen later. Therefore, all nonsingleton OOV words in lectDev and lectOther are added to the vocabulary resulting in a decrease of the OOV rate on lectEval to 0.49%.

5.1.4 Recognition results

Table 9 summarizes and compares the final recognition results using the techniques described above. The baseline is formed by using a speaker-independently trained acoustic model and no adaptation at all. With online adaptation using VTLN and cMLLR during decoding with the same acoustic model, the WER is reduced significantly by 21.1% on lectEval and 12.8% on lectDev. By using supervised speaker-adapted acoustic models instead the WER drops by about 11% on both sets. Using an adapted language model computed with an expanded vocabulary further decreases the WER by about 10–12% to a final WER of 11.2% on lectDev and 12.6% on lectEval.

5.2 Topic adaptation for MT

Adaptation of the MT component of our EPPS translation system towards the more conversational style of lectures and known topics 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 (see Table 11).

Table 10 Comparison of the perplexities for differently adapted target-language models

	lectDev	lectEval	Weight (%)
EPPS-T	631	1019	0.9
WEB01, common speech Web data	342	554	6.7
WEB02, topic related Web data	217	544	57.2
Lectures		711	35.1
Interpolated		348	

Table 11 Translation model and language model adaptation

	Unadapted		Target model		+ Language model	
	NIST	Bleu	NIST	Bleu	NIST	Bleu
t036+, text input	5.72	23.4	5.89	25.0	6.33	28.4
t036+, ASR input	5.06	17.9	5.15	18.9	5.60	23.0
l043, text input	5.27	19.6	5.35	20.3	5.48	21.6
l043, ASR input	4.80	16.6	4.85	16.9	5.05	19.0

Second, for (target) language model adaptation, Web data was collected in the same manner as described for the `RT-06S` ASR evaluation system. As can be seen in Table 10 the baseline `EPPS-T` language model based on EPPS text data has a very high perplexity. Therefore, two different Web data collections were performed. To adapt the language model to the more conversational style of lectures, common spontaneous speech expressions were extracted from some speech transcripts and used as queries. The resulting corpus `WEB01` had a size of about 90M words. A second collection was performed by using topic related n-grams as queries, whereby the n-grams were extracted from the development data. This results in a corpus `WEB02` of about 45M words. As can be seen from Table 10, both corpora got significantly lower perplexities than compared to `EPPS-T`. The reason for the lower perplexity for `WEB02` on `lectDev` is due to the queries which were extracted from the same data. While `WEB02` was taken as it is, `WEB01` was in addition filtered using a perplexity-based criterion on a document level. The lecture (`lectDev`) data itself achieved a perplexity of 711 on `lectEval`. From the interpolation weights given in Table 10, which were tuned on some other held-out data, it can be observed that the relevance of the `WEB02` data is much higher than the `WEB01` data. This might be due to some overlap between both corpora together with the fact that the second corpus is much more specific due to the topic n-grams. Overall a perplexity of 348 on the evaluation data is achieved.

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 Table 11 for English–Spanish. 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 improve performance significantly, especially when amplified by using language model adaptation with similar Web data.

6 Translatable speech segments

In speech translation systems the combination of ASR and MT is not always straightforward for optimal performance. In addition to the errors committed by the speech recognition leading to additional errors in the MT, the ASR hypotheses have to be resegmented such that the performance of the MT does not suffer because of them. Since almost all SMT systems are trained on data split at sentence boundaries this is commonly done by resegmenting the hypotheses according to automatically detected sentence boundaries.

Punctuation annotation is usually done by combining lexical and prosodic features (Liu 2004), whereas the combination is often done with the help of maximum entropy models (Huang and Zweig 2002) or CART-style decision trees (Kim and Woodland 2001).

For simultaneous translation systems, chunking of ASR hypotheses into useful translatable segments is even more critical and difficult (Fügen et al. 2006b). Due to the resulting latency, a global optimization over several ASR hypotheses as suggested in Matusov et al. (2006) is impossible. Instead a latency as short as possible, as explained in Sect. 2.2.1 is desirable. For English, as will be seen later, relatively good results were achieved with a latency of about three seconds corresponding to about 12 words at an average speaking rate.

This section presents and extends the work which the authors have published in Fügen and Kolss (2007). The work based on the experiments done by Cettolo and Federico (2006) in which the impact on translation performance of different text segmentation criteria was investigated. We address the questions on how chunking of ASR hypotheses as well as ASR reference transcripts into translatable segments, usually smaller then sentences, influence MT performance in an empirical study. Therefore, we compare different segmentation strategies on ASR hypotheses as well as on the reference transcripts. To measure the usefulness for simultaneous translation, we set the MT performance in relation to the average segment length and its standard deviation.

Simultaneously with Fügen and Kolss (2007), another work was published by Rao et al. (2007). This work aimed to improve the translation performance by optimizing the translation segment length only, instead of reducing the segment length to decrease translation system latency. The interesting result of this work is that the best performance is achieved with segments smaller than sentences, similar to the observations described here.

In Fig. 7 the output of the ASR system is a continuous stream of words and noises, and, based on this information, a segmentation algorithm has to find appropriate boundaries. Repetitions and other ungrammaticalities of spontaneous speech makes this a difficult task. The result of the best performing algorithm is shown as well, and, as can be seen, it is not perfect.

Section 6.1 presents the test data and systems used for these experiments, which differ from those introduced in Sect. 4, only in the translation direction. The reason for this was practical: at the time these studies were made, the Spanish–English translation system achieved the best translation performance.

ASR output:

```
so this is a really <breath> hot <breath> is it <breath> is
a dilemma <laugh> how do we provide access to global markets
and opportunities of one side and how to <breath> do we on the
other side <breath> do that while maintaining cultural diversity
while <mumble> while maintaining differences in languages and
differences between people <breath> so can technology provide
solutions for this <breath> guy <breath> that's what we're
ascension devoted to you <breath> at our laboratory building
technologies that provide tools for people <breath> to <breath>
communicate <breath> <breath> the goal for good building is to
<breath> build <breath> <mumble>
```

Segmentation using the proposed algorithm:

```
so this is a really . hot . is it is a dilemma . how do we provide
access to global markets and opportunities of one side and . how
to do we on the other side do that . while maintaining cultural
diversity while while maintaining differences in languages and
differences between . people so can technology provide solutions
for this guy . that is what we are ascension devoted to you at our
laboratory building technologies that . provide tools for people
to communicate . the goal for good building is to build .
```

Fig. 7 Example of an unsegmented output of the ASR system and the segmented output using the best performing algorithm. Segment boundaries are marked with periods. Recognized human and nonhuman noises are written in brackets

6.1 Data and systems

At the time of the experiments only a Spanish–English translation system was available (Kolss et al. 2006), which was trained using minimum error rate training (Och 2003) on the same parallel EPPS data as that mentioned in Sect. 4.3, but with the translation direction from Spanish to English. It is later shown in Sect. 7 that the results obtained are also transferable to the other translation direction. For decoding, a word reordering window size of four was used in our experiments.

As test data, we selected the 2006 Spanish–English TC-STAR development data consisting of three hours (14 sessions) of nonnative Spanish speech recorded at the European Parliament. We used ASR hypotheses as well as reference transcripts for the experiments, whereas the Spanish hypotheses were generated with a system trained within TC-STAR on Parliament plenary sessions (Stüker et al. 2007; Paulik and Waibel 2008). The case-insensitive WER was 8.4%.

6.2 Experimental results and discussion

6.2.1 Scoring MT with different segmentations

The commonly used metrics for the automatic evaluation of MT output, such as the Bleu (Papineni et al. 2002) and NIST (Doddington 2002) metrics, were originally developed for translation of written text, where the input segment boundaries correspond to the reference sentence boundaries. This is not the case for translation of spoken language where the correct segmentation into sentence-like units is unknown and must be produced automatically by the system.

In order to be able to use the established evaluation measures, the translation output of the automatically produced segments must be mapped to the reference translation segments in advance of the scoring procedure. This is done by using the method described in Matusov et al. (2005), which takes advantage of the edit distance algorithm to produce an optimal resegmentation of the hypotheses for scoring which is invariant to the segmentation used by the translation component.

The Bleu scores for the end-to-end evaluation were obtained by using this method using two reference translations. Since the alignment was performed on a per-session level the result is invariant in relation to the number of segments produced for a single session. The scoring was done case-insensitively without taking punctuation marks into account.

6.2.2 Baselines

Resegmenting ASR hypotheses at sentence boundaries for MT is the most common approach for speech translation systems. For this reason, the translation scores obtained by translating ASR hypotheses as well as reference transcripts split at sentence boundaries (`sent`) serve as one baseline for the following experiments. As can be seen in Table 12, we obtained a Bleu score of 36.6% by translating ASR reference transcripts and a score of 33.4% for ASR hypotheses, which clearly shows the influence of the ASR performance on MT quality. The average segment length was 30 words with a standard deviation of 22.

Another baseline is obtained by taking all punctuation marks as split points (`punct`). By doing this, the average segment length could be reduced to an acceptable nine words with almost no decrease in the translation score. The reason for this is that punctuation marks are usually used to represent semantic boundaries.

6.2.3 Destroying the semantic context

In this section, we analyzed how MT performance is affected by destroying the semantic context of an utterance independently of the applicability for simultaneous translation. Following the experiments in Cettolo and Federico (2006) we simply cut the merged utterances of a single session every n word (`fixed`). The results are given in Table 12 for $n \in \{7, 11, 15\}$ and provide a lower bound for the chunking strategies that will be presented below. As expected, the decrease in segment size, i.e. the destruction of the semantic context, affected the translation scores significantly.

6.2.4 Using acoustic features

Following the finding of Chen (1999), that pauses closely correspond to punctuation, we used the information about nonspeech regions in the ASR hypotheses for resegmentation. For nonspeech regions, we used recognized silences and nonhuman

Table 12 Bleu scores obtained on ASR reference transcripts (Ref) as well as on ASR hypotheses (ASR) together with the average (avg) segment length and standard deviation (sd)

Chunking strategy	Segment length		Correlation		Bleu	
	avg	sd	Precision	Recall	Ref	ASR
Baseline						
sent	30.1	22.1	98.5	26.5	36.59	33.41
punct	9.2	6.7	100.0	98.6	35.91	33.31
Length based						
fixed-7	7.0	0.2	22.8	26.5	30.13	27.50
fixed-11	11.0	0.7	22.6	16.8	32.07	29.53
fixed-15	15.0	0.6	23.8	13.0	33.64	30.66
sent-0.5	15.3	11.3	59.0	31.8	35.08	
sent-0.25	10.3	8.5	44.9	36.1	33.67	
Using acoustic features						
pause-0.1	8.2	5.7	59.3	58.3		31.86
pause-0.2	12.1	11.0	66.3	44.5		32.53
pause-0.3	17.0	19.6	71.5	34.0		32.62
pause-var15	7.5	3.1	59.4	61.4		31.34
pause-var20	9.8	4.3	64.6	53.0		31.87
pause-var25	11.8	5.3	68.3	46.9		32.36
asr	13.4	9.0	70.1	42.6		32.68
Using also other features						
lm	10.9	9.8	73.1	54.1		32.90
lm-10	8.7	5.6	67.3	62.4		32.36

Precision and recall of segmentation boundaries to punctuation marks are given as well

noises, whereas successive noises and silences were merged together. For the translation scores (`pause`) in Table 12 we used different nonspeech duration thresholds (0.1, 0.2, and 0.3 seconds). As expected, the results are significantly better than those obtained with the chunking strategies in Sect. 6.2.3.

The standard deviations of the segment lengths achieved with the nonspeech chunking strategies are still too large for use in simultaneous translation. By splitting the ASR hypotheses at the longest nonspeech interval within a region of a maximum number of words (`var15`, `var20`, `var25`, with chunks of maximally 15, 20, 25 words), the standard deviation could be significantly reduced without decreasing the translation quality when comparing fixed and variable nonspeech chunking strategies having a similar average segment length.

Overall, chunking strategies using nonspeech regions are simple and require no additional context information, but nonetheless, achieving relatively good translation scores. While precision and recall show a good correlation between nonspeech regions and punctuation marks, only a slight correlation between precision and Bleu score could be observed. This let us come to the conclusion that an optimal chunking

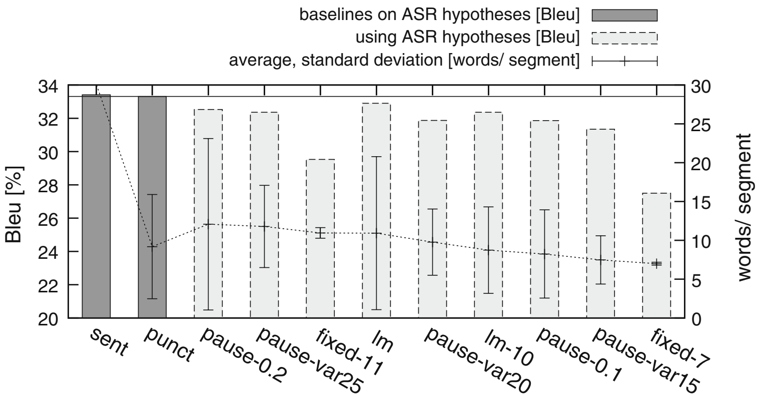

Fig. 8 Results obtained by chunking ASR hypotheses sorted descending according to their average segment length. Bleu scores (*left axis*) are represented by boxes, while the markers in the middle of the boxes give the average segment length (*right axis*) together with the standard deviation

strategy for MT does not necessarily have to have a high correlation with punctuation marks. Instead, additional features have to be considered as well.

6.2.5 Using other features

In Cettolo and Falavigna (1998), prosodic as well as lexical information (punctuation marks, filled pauses and human noises) was used to detect semantic boundaries automatically. According to Cettolo and Falavigna, a trigram language model was trained, whereas all punctuation marks in the training corpus were substituted by a boundary tag BD. But instead of doing a global optimization over the whole sentence, the decision was made using local information only, i.e. (a) by setting the language model probabilities $Pr(w_{i-1}, w_i, \text{BD}, w_{i+1}, w_{i+2})$ and $Pr(w_{i-1}, w_i, w_{i+1}, w_{i+2})$ in relation to each other and (b) by using additional thresholds for the nonspeech gap in between w_i and w_{i+1}. As can be seen in Table 12, (lm) this chunking strategy outperforms all other strategies using acoustic features only. A precision of 73% and a recall of 54% was measured. Nevertheless a standard deviation of 10 from the average of 11 words was measured. Therefore, in a second experiment we relaxed the above-mentioned thresholds when no split point was found after ten words (lm-10). By doing this, the standard deviation could be almost halved with only a minor decrease in Bleu score.

Overall, it can be seen in Fig. 8 that sentence boundaries are a good criterion for utterance chunking, but are inapplicable for simultaneous translation because of the high average sentence length. Chunking strategies based on nonspeech regions are simple and require no additional context information, but nonetheless achieve relatively good translation scores. Nevertheless, a more sophisticated approach using lexical features provided by a language model in addition outperforms all chunking strategies using acoustic features only.

Overall, the chunking strategy lm-10 outperforms the manual sentence-based segmentation and performs almost identically to the manual segmentation using

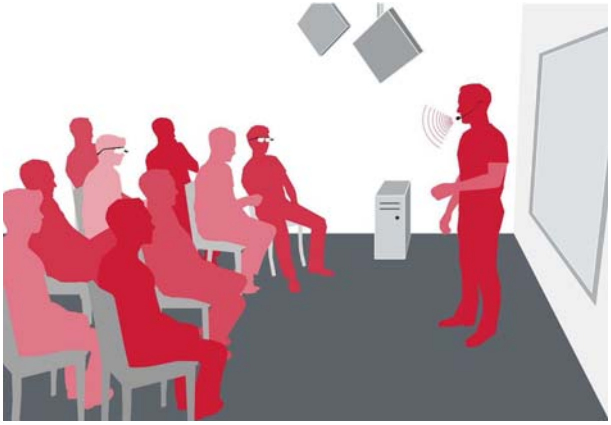

Fig. 9 Configuration of the lecture translation system. Ceiling-mounted targeted audios and heads-up display goggles are used to deliver translations to individual people in the audience only

punctuation marks `punct`. Furthermore, it has the desired average segment length of nine and a small standard deviation of six and is therefore well suited for simultaneous translation systems.

7 A system for simultaneous speech translation

Given the description of the single components of the simultaneous translation system in the previous section, the configuration of the lecture translation system, shown in Fig. 9, is explained in more detail. Furthermore, some latency and real-time issues are discussed.

7.1 System overview

Figure 2 already showed the schematic overview of the architecture of the simultaneous translation system and its information flow. All components of the simultaneous translation system are connected together in a client-server framework, similar to the one described in Fügen et al. (2001). It is designed in a way where several consumers can be used to process the output of a single or multiple producers to reduce the latency of the system, if necessary. If required, the architecture also allows for combining for example the hypotheses of several ASR components on the same input signal to improve the recognition quality by easily adding another consumer responsible for the combination. The only requirement in this case is that the latency of the system be not affected.

When offering simultaneous translation services there is always the question on how to deliver the translation to an audience. In the case of human interpreters the translation is delivered in spoken form for which headphones are typically used. An automatic system has the advantage that the translation can be delivered as well in written form. In the following, the advantages and disadvantages of the different transmission forms, types, and output devices are discussed in more detail.

- *In written form*: One of the most prominent transmission types is presenting the translation as subtitles on a projection screen. The advantages of this transmission type are that it is easy to implement, does not require any additional hardware and can be presented to several people at the same time. However, with the necessity to display an increasing number of output languages at the same time, clarity of the presentation also suffers. In contrast to this, with display devices assigned to a single person or a small group of persons of the same mother tongue, it is possible to deliver the required target-oriented translation language only, but then it is necessary to purchase additional hardware. The additional costs could be reduced significantly by also supporting standard handheld devices like PDAs or cellphones, which are becoming nowadays more widely used. All above-mentioned output devices have the major disadvantage that for reading the translation, one is unable to look at the lecturer or presentation slides at the same time. For this reason, heads-up display goggles (see Fig. 10a) might be an ideal solution, because they enable the wearer to see both the lecturer and translation at the same time. However, it is very expensive to equip an audience with such a device.
- *In spoken form*: The other most prominent transmission form is delivering the translation with the help of a speech synthesis in spoken form. Using standard loudspeakers for this is only suitable when the whole audience has the same mother tongue. Otherwise, the different spoken translations and original speech would interfere with each other. Using headphones solves this problem, but requires additional equipment for the audience. Furthermore, it hinders the audience from communicating with each other, because it becomes difficult to speak with a neighbor without taking off the headphones. For this reason another solution came to mind: beam steered sound beams or "target audio devices", which allow delivery of the translation to a desired group of people located in the beam using a single device only, without disturbing other people in the audience located outside of the beam. With this solution, communication between people in the audience is not hindered and the audience is not forced to wear additional personal devices. Furthermore, due to the fact that several people can be addressed with a single device, the additional cost can also be reduced. Figure 10b shows such a targeted audio device: a parametric array consisting of several ultrasound loudspeakers, which was proposed by Olszewski et al. (2005).

Although several problems could be solved with the help of these devices, one problem still exists, which we call the "delivery rate problem". To keep the latency of the system constant, the system has to consume input as fast as it is produced and therefore has to deliver output at the same rate. This is especially problematic for language pairs for which an expression with the same meaning consists of more words in the target language than in the source language, like for example with English and Finnish. Furthermore, the automatic system is not perfect, which means that understanding an imperfect written or spoken translation is not an easy task, but increasing the delivery rate of the translation to keep up with the lecturer will make this task even more difficult.

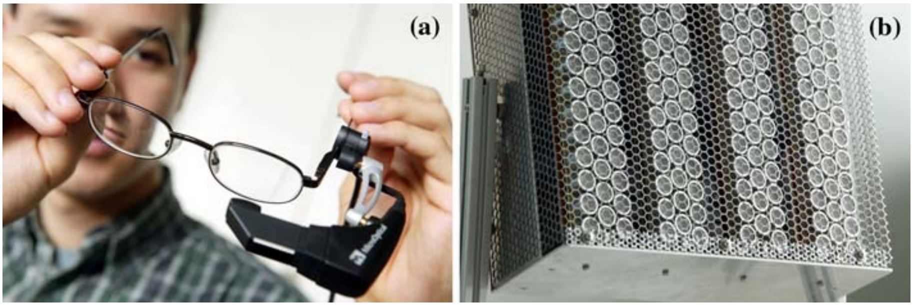

Fig. 10 Different output devices: **a** Translation goggles. **b** The targeted audio device

8 Latency and real-time issues

Besides the latency, which is caused by the serialization of the components of the simultaneous translation system, i.e. ASR, resegmentation and MT, the RTF of the components themselves for processing a given input plays an important role, in which the influence of the speech recognition is the largest.

The most common strategy for reducing the latency in a speech recognition system is to start with the decoding as soon as enough recorded data is available, i.e. without waiting until the recording has finished. The most prominent example for this strategy, called "run-on decoding", is a commercial dictation system. Typically, voice activity detection (VAD) is used in advance of decoding to prevent the speech recognition from decoding huge amounts of silence or background noise, which often cause recognition errors. However, for lectures there is typically no speaker change and also a more or less continuous stream of speech, so that VAD is less important. Furthermore, an additional latency is introduced by VAD, because several consecutive frames normally have to be classified as speech to start a new segment and because of the algorithm itself. Therefore, a relatively simple, energy-based VAD is used, which is responsible for a rough speech/nonspeech classification only.

To decrease the latency further, the speech recognizer was modified to deliver partial, single-best hypotheses even during the decoding of a segment. Our implementation returns only those hypotheses which, independent of the search-tree pruning, will not change in the future. Each partial hypothesis is directly passed to the resegmentation component. The disadvantage of this strategy with our Ibis decoder is that an additional language model rescoring pass is no longer possible.

As can be seen in Table 13, while a smaller interval length produces a smaller latency, the RTF increases slightly due to the overhead for processing more segments. Latencies achieved for the systems using a 1-second interval are significantly lower than those obtained with a 2-second interval if the total RTF is lower than 1. When applying incremental adaptation, it can be seen that the RTF increases due to the overhead for computing the adaptation parameters, but the latency is reduced, because the better adapted acoustic models are leading to a smaller search space and a faster decoding speed. In the case of the system with incremental adaptation using a 1-second interval, it can be seen what happens if the total RTF is higher than 1.0. The system cannot

Table 13 Comparison of the WER, latency in seconds, RTF of the decoding itself and the total RTF for different audio interval durations (in seconds) on LectDev

Interval	WER (%)	Latency	RTF
Baseline	16.9		1.10
0.5	17.3	3.48	1.05
1.0	16.7	3.60	0.98
2.0	16.6	5.10	0.96
3.0	16.4	6.83	0.95
5.0	16.6	10.36	0.93
Baseline adapted	15.1		1.20
+ Gaussian selection	15.4		0.70
1.0 Adapted	15.3	6.33	1.18
+ Gaussian selection	15.5	2.16	0.88
2.0 Adapted	15.3	5.80	1.08
+ Gaussian selection	15.2	3.61	0.78

keep up with the speaking rate of the input speech leading to high latency. In such cases, applying additional speed-up techniques such as Gaussian selection is essential.

As mentioned above, no lattice rescoring was used in all cases. The respective WERs for the two baselines with lattice rescoring can be found in Table 9: an absolute difference of 0.5–0.8 can be observed. Figure 11 shows an example of the output of the ASR system before and after language model rescoring. As can be seen, due to a higher weighting of the language model scores compared to the acoustic model scores, the influence of the language model is increased which results a more accurate hypothesis.

For translation, a reordering window of four words was used, resulting in a translation speed of about 50 words per second. That means that the translation of a segment with an average length of ten words does not take longer than speaking a single word.

To reach the above-mentioned real-time and latency restrictions, the simultaneous translation system has to run on a multiprocessor workstation. If it is required to run the system on a laptop, the processing speed of the components has to be increased even more. For the speech recognizer this can be done by reducing the search space pruning thresholds or by applying Gaussian selection (Fritsch and Rogina 1996). For MT, the reordering window can be reduced to a size of two.

9 End-to-end evaluation results

The evaluation of the final end-to-end system for the English–Spanish direction was carried out on six excerpts, three each from the two talks t036+ and l043 as described in Sect. 4.1.4. The excerpts focus on a different topic each and represent a total of about 36 minutes and 7,200 words. The WER of the ASR component on this subset was 11.9% with an OOV rate of 0.5%. For computing the MT scores, the two available reference translations were automatically resegmented using the approach described in Matusov et al. (2005). Table 14 shows the automatic evaluation results of the trans-

Reference: o. k. so if i have a perception here and i am getting a classification of one here and i do not want that i compute the distance between the output that my random initial perceptron gives me and the desired output that i would like to have and i compute then the derivative between the of the error with respect to the weight o. k. i just mathematically compute a derivative

Before lattice rescoring: o. k. ** ** * IS a JAVA PERCEPTRON here and * AND getting a classification of one here and i do not want that i compute the distance between the output that my RENDERED AN initial PERCEPTION VIEWS FOR me and the desired output that THEY would like to HALF and i compute THAN the DO RELATIVE between the of the error with respect to the weight o. k. i just YEAH PHONETICALLY compute THE DERIVATIVES

After lattice rescoring: o. k. so ** * THEY ARE PERCEPTRON here and * AND getting a classification of one here and i do not want that i compute the distance between the output that my random initial PERCEPTION gives me and the desired output that THEY would like to have and i compute then the derivative between the of the error with respect to the weight o. k. i just DEFINITELY compute THE DERIVATIVES

Fig. 11 Example of an unsegmented output of the ASR system before and after lattice rescoring. Capitalized words are substitution or insertion errors and stars are deletion errors

Table 14 ASR and SMT automatic evaluation results

	ASR WER (%)	SMT Bleu	SMT NIST
EPPS (TC-STAR 2007)	6.9	40.60	9.19
Lectures (our system)	11.9	28.94	7.07

lated output for the simultaneous translation system on `Lectures` and compares them with the results of the final TC-STAR 2007 evaluation system for parliamentary speeches `EPPS` (see Hamon et al. 2007).

9.1 Methodology

The human evaluation was carried out using the same methodology as used in the TC-STAR project (Hamon et al. 2007). A professional human interpreter was asked to translate the same six excerpts, and the outputs of the automatic system and the human interpreter were then presented to human judges. Of the 20 judges, all native Spanish speakers, half of them did not have specific knowledge of the technical domain of the lectures and no specific experience, while the other judges had knowledge of the domain. Each judge evaluated either the automatic excerpts or the human excerpts, and each segment was thus evaluated by 10 different judges.

9.1.1 Evaluation protocol

Excerpts were presented to the judges via an online interface. For each excerpt, judges were instructed to:

- listen to the current excerpt only once;
- answer a comprehension questionnaire;
- evaluate the fluency;
- go to the next excerpt.

For the fluency evaluation, evaluators were asked to judge the quality of the language on a 5-point scale, with a value of 1 for lowest and a value of 5 for highest quality.

9.1.2 Comprehension questionnaire

After listening to an excerpt, judges had to fill out a comprehension questionnaire of ten questions, five general questions and five more specific ones, for a total of 60 questions. Three types of questions were asked:

- simple factual (70%),
- yes/no (20%),
- list (10%).

This was done in order to diversify the difficulty of the questions and test the system at different levels. After all the evaluations were done, an objective verification of the answers, or validation, was performed by a Spanish native speaker, comparing the judges' answers to the reference answers. Another objective validation was also done on the SLT and ASR parts in order to find where the information was lost, as well as on the interpreter output.

Two kinds of criteria were used, the first with four values, for analysis and study, the second with a correspondence with binary values, in order to present the results in a clearer way. These criteria were used for all kinds of validations. The four values are:

- 0: the information or answer is totally missing;
- 1: the information or answer is partially correct, but clearly incomplete;
- 2: the information or answer is really close or with some more information, but which does not correspond exactly to the reference answer;
- 3: the information or answer is undeniably correct.

The binary values simplify the scale according to:

- for 0 and 1 = 0 : the information or answer is incorrect;
- for 2 and 3 = 1 : the information is correct.

9.2 Results

We compared the results of the human interpreter and the automatic system on the lecture task with the results obtained by the TC-STAR system on the EPPS task under otherwise identical conditions.

Table 15 Human evaluation, overall fluency: score between 1 (worst) and 5 (best)

	Interpreter	Automatic system
EPPS	4.03	2.05
Lectures	3.03	2.35

Table 16 Human evaluation, overall comprehension, binary classification: figures show percentage of correct answers

	Interpreter	Automatic system
EPPS	74	64
Lectures	74	52

9.2.1 Fluency

Table 15 shows that while the fluency of the automatic system was judged to be slightly better than the TC-STAR system, the fluency of human interpreters is worse on the lectures than on the EPPS task. This is most likely due to the more spontaneous speech style of lectures and a higher familiarity of interpreters with the topics and the typical expressions used in the European Parliament.

For the human interpreter, the reasons for fluency errors are the lengthening of words, hesitations, pauses between syllables and catching breath, other careless mistakes as well as self-corrections of wrong interpreting. For the automatic system, the performance of the translation component contributes more to the fluency errors than the performance of the ASR component. Some English words are not translated at all and the word ordering (especially for compound words) is not always correct.

9.2.2 Comprehension

Table 16 shows the percentage of correct answers given by judges from the output of interpreters and the automatic systems. The automatic systems perform worse than the interpreters, and our system is worse than the TC-STAR system. This is clearly in accordance with the difference in WER and Bleu between the EPPS and lecture domains.

In addition, an objective verification of the presence of the information needed to answer the questions was performed by a native Spanish speaker. Table 17 shows the percentage of answers present or the maximal answers found in the Spanish translation output and in the intermediate outputs of the ASR and MT components. The comparison shows that for the lecture task, the human interpreter was able to retain 85% of the information needed to answer all questions, while the judges, who could listen to the interpreter output only once, then could answer 74% of the questions. The automatic system presented 58% of the necessary information to the judges who then answered 52% of the questions correctly; the ASR component already loses 17% of the information, compared to 3% for the TC-STAR system.

To compare the translation quality of the automatic system with the quality of the human interpreter objectively, the results were limited to the questions for which the answers were included in the interpreter files, as shown in Table 18.

Table 17 Human evaluation, objective comprehension: percentage of answers present in the output

	Interpreter	Automatic system	MT	ASR
EPPS	91	89	92	97
Lectures	85	58	65	83

Table 18 Human evaluation, comprehension of results limited to questions for which the answers were included in the interpreter files: percentage of correct answers

	Interpreter	Automatic system	TTS	MT	ASR
EPPS	80	66	91	93	97
Lectures	80	53	60	70	80

Here it can be seen that due to the limitation, not only are the interpreters getting better (from 74% to 80%), but also the automatic systems are getting slightly better (from 64% to 66% and from 52% to 53%). But again, there is already a loss of 20% in the ASR.

10 Conclusion

In this paper, we have presented a first real-time simultaneous speech translation system for spontaneous technical lectures. The system presented borrows heavily from ongoing advances in domain-unlimited speech translation technologies and extends these along five important additional dimensions:

1. the real-time requirement of simultaneous interpretation,
2. the disfluent nature of spontaneous (unprepared, not read speech) monologues or lectures,
3. special terms, named entities, special expressions found in lectures on specialized topics,
4. the problem of online sentence segmentation in real time (what is a sentence?), and
5. the human factors of simultaneous interpretation, or how to deliver translation output unobtrusively during an ongoing lecture.

As seen, adaptation is of utmost importance, but still takes considerable effort. To make the system more flexible the automatic harvesting of related data from the Web for topic adaptation has to be extended. Together with speaker profiles, the adapted systems can be easily re-used as a basis for subsequent talks and further adaptation.

Although not perfect, an approach to resegmenting the ASR output was presented. But the problem is of principle nature; only the MT system knows if a segment presents a semantic unit ideal for translation or if it is better to wait for more input. Therefore, future research should go into the direction of supporting the streaming approach as used in speech recognition also in MT. This moves the segmentation into the MT system.

Extensive user experimentation appears to suggest that MT systems are already achieving usable results. More interestingly, machine interpreters seem to be held back only by ASR or MT component deficiencies, deficiencies that could conceivably be improved with continuing advances and longer-term data collection. Human translators, by contrast, appear to be held back predominantly by the cognitive limitations of doing this challenging task under the pressure of the real-time requirement drawing on a limited human short-term memory, and these limitations are ones that cannot easily be resolved. These intriguing results leave room for speculation on whether superhuman simultaneous interpretation performance may one day be possible by machine.

Acknowledgements Such a complex system is typically not the work of only a few people, wherefore we would like to thank Susi Burger for her help in transcribing and translating the lecture data together with her students, Matthias Paulik for his help in improving the target-language model by collecting some Web data, for providing the Spanish ASR system used for the experiments described in Sect. 6, and for his support in showing this system to guests at our lab at the Carnegie Mellon University, Sebastian Stüker for his support in preparing several other demonstrations of the system at Universität Karlsruhe (TH) and at conferences, and Stephan Vogel for his support in MT in general. Furthermore, we would like to thank Olivier Hamon, Djamel Mostefa, and Khalid Choukri from ELDA (Evaluations and Language resources Distribution Agency) for their support with the human evaluation.

References

Accipio Consulting (2006) Sprachtechnologien für Europa [Language technologies for Europe]. ITC IRST, Trento, Italy. Available at http://www.tc-star.org/pubblicazioni/D17_HLT_DE.pdf. Accessed 29 Oct 2008

Al-Khanji R, El-Shiyab S, Hussein R (2000) On the use of compensatory strategies in simultaneous interpretation. Meta J Traduc 45:544–557

Atal B (1974) Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification. J Acoust Soc Am 55:1304–1312

Bahl L, Brown P, de Souza P, Mercer R (1986) Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: ICASSP ’86, IEEE international conference on acoustics, speech, and signal processing, Tokyo, Japan, pp 49–52

Bain K, Basson S, Faisman A, Kanevsky D (2005) Accessibility, transcription, and access everywhere. IBM Syst J 44:589–603

Barik HC (1969) A study of simultaneous interpretation. PhD thesis, University of North Carolina at Chapel Hill

Bellegarda JR (2004) Statistical language model adaptation: review and persepectives. Speech Commun 42:93–108

Black AW, Taylor PA (1997) The Festival speech synthesis system: system documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Edinburgh, Scotland

Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1994) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19:263–311

Bulyko I, Ostendorf M, Stolcke A (2003) Getting more mileage from Web text sources for conversational speech language modeling using class-dependent mixtures. In: HLT-NAACL 2003 Human language technology conference of the North American chapter of the Association for Computational Linguistics, Companion volume: short papers, student research workshop, demonstrations, tutorial abstracts, Edmonton, Alberta, Canada, pp 7–9

Burger S, MacLaren V, Waibel A (2004) ISL meeting speech part 1, catalog nbr LDC2004S05, Linguistic Data Consortium, Philadelphia, PA

Carletta J, Ashby S, Bourban S, Flynn M, Guillemot M, Hain T, Kadlec J, Karaiskos V, Kraaij W, Kronenthal M, Lathoud G, Lincoln M, Lisowska A, McCowan I, Post W, Reidsma D, Wellner P (2005) The AMI meeting corpus: A pre-announcement. In: 2nd joint workshop on multimodal interaction and related machine learning algorithms MLMI 05, Edinburgh, Scotland, pp 28–39

Cettolo M, Falavigna D (1998) Automatic detection of semantic boundaries based on acoustic and lexical knowledge. In: Fifth international conference on spoken language processing, ICSLP'98, Sydney, Australia, pp 1551–1554

Cettolo M, Federico M (2006) Text segmentation criteria for statistical machine translation. In: Salakoski T, Ginter F, Pyysalo S, Pahikkala T (eds) Advances in natural language processing, 5th international conference, FinTAL 2006, Turku, Finland, LNCS 4139. Springer Verlag, Berlin, pp 664–673

Cettolo M, Brugnara F, Federico M (2004) Advances in the automatic transcription of lectures. In: ICASSP 2004, IEEE international conference on acoustics, speech, and signal processing, Montreal, Canada, pp 769–772

Chen CJ (1999) Speech recognition with automatic punctuation. In: Sixth European conference on speech communication and technology (Eurospeech'99), Budapest, Hungary, pp 447–450

de Mori R, Federico M (1999) Language model adaptation. In: Ponting K (ed) Computational models of speech pattern processing. Springer Verlag, Berlin, pp 280–303

Doddington G (2002) Automatic evaluation of MT quality using n-gram co-occurrence statistics. In: Proceedings of human language technology conference 2002, San Diego, CA, 138–145

Eide E, Gish H (1996) A paramteric approach to vocal tract length normalization. In: 1996 IEEE international conference on acoustics, speech, and signal processing, Atlanta, Georgia, pp 346–348

Finke M, Geutner P, Hild H, Kemp T, Ries K, Westphal M (1997) The Karlsruhe-VERBMOBIL speech recognition engine. In: 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP'97), Munich, Germany, pp 83–86

Fiscus J (1997) A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER). In: Proceedings of the 1997 IEEE workshop on automatic speech recognition and understanding, Santa Barbara, CA, pp 347–352

Fiscus J, Garofolo J, Przybocki M, Fisher W, Pallett D (1998) 1997 English broadcast news speech (HUB4), catalog nbr LDC98S71, Linguistic Data Consortium, Philadelphia, PA

Foster G, Kuhn R, Johnson H (2006) Phrasetable smoothing for statistical machine translation. In: EMNLP 2006 conference on empirical methods in natural language processing, Sydney, Australia, pp 53–61

Fritsch J, Rogina I (1996) The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture Gaussians. In: 1996 IEEE international conference on acoustics, speech, and signal processing, Atlanta, Georgia, pp 837–840

Fügen C, Kolss M (2007) The influence of utterance chunking on machine translation performance. In: Interspeech 2007, 8th annual conference of the International Speech Communication Association, Antwerp, Belgium, pp 2837–2840

Fügen C, Westphal M, Schneider M, Schultz T, Waibel A (2001) LingWear: a mobile tourist information system. In: Proceedings of the first international conference on human language technology research, San Diego, California, 5 pp

Fügen C, Ikbal S, Kraft F, Kumatani K, Laskowski K, McDonough JW, Ostendorf M, Stüker S, Wölfel M (2006a) The ISL RT-06S speech-to-text system. In: Renals et al (2006), pp 407–418

Fügen C, Kolss M, Paulik M, Waibel A (2006b) Open domain speech translation: from seminars and speeches to lectures. In: TC-STAR workshop on speech to speech translation, Barcelona, Spain, pp 81–86

Furui S (1986) Cepstral analysis technique for automatic speaker verification. IEEE T Acoust Speech Signal Proc 34:52–59

Furui S (2005) Recent progress in corpus-based spontaneous speech recognition. IEICE T Inform Syst E88-D:366–375

Furui S (2007) Recent advances in automatic speech summarization. In: Symposium on large-scale knowledge resources (LKR 2007), Tokyo, Japan, pp 49–54

Gales MJF (1998) Maximum likelihood linear transformations for HMM-based speech recognition. Comput Speech Lang 12:75–98

Garofolo JS, Michel M, Stanford VM, Tabassi E, Fiscus J, Laprun CD, Pratz N, Lard J (2004) NIST meeting pilot corpus speech, catalog nbr LDC2004S09, Linguistic Data Consortium, Philadelphia, PA

Geutner P, Finke M, Scheytt P (1998) Adaptive vocabularies for transcribing multilingual broadcast news. In: Proceedings of the 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP '97), Seattle, Washington, pp 925–928

Glass J, Hazen TJ, Cyphers S, Malioutov I, Huynh D, Barzilay R (2007) Recent progress in the MIT spoken lecture processing project. In: Interspeech 2007, 8th annual conference of the International Speech Communication Association, Antwerp, Belgium, pp 2553–2556

Godfrey JJ, Holliman E (1993) Switchboard-1 transcripts, catalog nbr LDC93T4, Linguistic Data Consortium, Philadelphia, PA

Gollan C, Bisani M, Kanthak S, Schlüter R, Ney H (2005) Cross domain automatic transcription on the TC-STAR EPPS corpus. In: ICASSP, 2005 IEEE conference on acoustics, speech, and signal processing, Philadelphia, PA, pp 825–828

Gollan C, Hahn S, Schlüter R, Ney H (2007) An improved method for unsupervised training of LVCSR systems. In: Interspeech 2007, 8th annual conference of the International Speech Communication Association, Antwerp, Belgium, pp 2101–2104

Graff D (1994) UN parallel text (complete), catalog nbr LDC94T4A, Linguistic Data Consortium, Philadelphia, PA

Graff D (2003) English gigaword, catalog nbr LDC2003T05, Linguistic Data Consortium, Philadelphia, PA

Graff D, Garofolo J, Fiscus J, Fisher W, Pallett D (1997) 1996 English broadcast news speech (HUB4), catalog nbr LDC97S44, Linguistic Data Consortium, Philadelphia, PA

Hamon O, Mostefa D, Choukri K (2007) End-to-end evaluation of a speech-to-speech translation system in TC-STAR. In: Machine translation summit XI, Copenhagen, Denmark, pp 223–230

Henderson JA (1982) Some psychological aspects of simultaneous interpreting. Incorp Ling 21(4):149–150

Hendricks PV (1971) Simultaneous interpreting: a practical book. Longman, London

Huang J, Zweig G (2002) Maximum entropy model for punctuation annotation from speech. In: 7th international conference on spoken language processing (ICSLP 2002, Interspeech 2002), Denver, Colorado, pp 917–920

Huang J, Westphal M, Chen SF, Siohan O, Povey D, Libal V, Soneiro A, Schulz H, Ross T, Potamianos G (2006) The IBM rich transcription spring 2006 speech-to-text system for lecture meetings. In: Renals et al (2006), pp 432–443

Janin A, Edwards J, Ellis D, Gelbart D, Morgan N, Peskin B, Pfau T, Shriberg E, Stolcke A, Wooters C (2004) ICSI meeting speech, catalog nbr LDC2004S02, Linguistic Data Consortium, Philadelphia, PA

Jones R (1998) Conference interpreting explained. St. Jerome Publishing, Manchester

Kim J-H, Woodland PC (2001) The use of prosody in a combined system for punctuation generation and speech recognition. In: Eurospeech 2001 Scandinavia, 7th European conference on speech communication and technology, 2nd Interspeech event, Aalborg, Denmark, pp 2757–2760

Klakow D, Peters J (2002) Testing the correlation of word error rate and perplexity. Speech Commun 38:19–28

Koehn P, Axelrod A, Mayne AB, Callison-Burch C, Osborne M, Talbot D (2005) Edinburgh system description for the 2005 IWSLT speech translation evaluation. In: Proceedings of international workshop on spoken language translation, Pittsburgh, PA

Kolss M, Zhao B, Vogel S, Hildebrand AS, Niehues J, Venugopal A, Zhang Y (2006) The ISL statistical machine translation system for the TC-STAR spring 2006 evaluation. In: TC-STAR workshop on speech to speech translation, Barcelona, Spain

Kopczynski A (1994) Bridging the gap: empirical research in simultaneous interpretation. John Benjamins, Amsterdam/Philadelphia

Lamel LF, Schiel F, Fourcin A, Mariani J, Tillmann HG (1994) The translanguage English database TED. In: Third international conference on spoken language processing (ICSLP 94), Yokohama, Japan, pp 1795–1798

Lamel L, Bilinski E, Adda G, Gauvain J-L, Schwenk H (2006) The LIMSI RT06s lecture transcription system. In: Renals et al. (2006), pp 457–468

Lamel L, Gauvain J-L, Adda G, Barras C, Bilinski E, Galibert O, Pujol A, Schwenk H, Zhu X (2007) The LIMSI 2006 TC-STAR EPPS transcription systems. In: ICASSP 2007, international conference on acoustics, speech, and signal processing, Honolulu, Hawaii, pp 997–1000

Lederer M (1978) Simultaneous interpretation: units of meaning and other features. In: Gerver D, Sinaiko HW (eds) Language interpretation and communication. Plenum Press, New York, pp 323–332

Leggetter CJ, Woodland PC (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang 9:171–185

Liu Y (2004) Structural event detection for rich transcription of speech. PhD thesis, Purdue University, West Lafayette, IN

Lööf J, Bisani M, Gollan C, Heigold G, Hoffmeister B, Plahl C, Schlüter R, Ney H (2006) The 2006 RWTH parliamentary speeches transcription system. In: Interspeech 2006 – ICSLP, ninth international conference on spoken language processing, Pittsburgh, PA, pp 105–108

Mani I (2001) Automatic summarization. John Benjamins, Amsterdam

Matusov E, Leusch G, Bender O, Ney H (2005) Evaluating machine translation output with automatic sentence segmentation. In: Proceedings of international workshop on spoken language translation, Pittsburgh, PA

Matusov E, Mauser A, Ney H (2006) Automatic sentence segmentation and punctuation prediction for spoken language translation. In: International workshop on spoken language translation, Kyoto, Japan, pp 158–165

Matusov E, Leusch G, Banchs RE, Bertoldi N, Déchelotte D, Federico M, Kolss M, Lee Y-S, Mariño JB, Paulik M, Roukos S, Schwenk H, Ney H (2008) System combination for machine translation of spoken and written language. IEEE T Audio Speech Lang Proc 16:1222–1237

Morimoto T, Takezawa T, Yato F, Sagayama S, Tashiro T, Nagata M, Kurematsu A (1993) ATR's speech translation system: ASURA. In: European conference on speech communication and technology 1993, Eurospeech 1993, Berlin, Germany, pp 1291–1294

Moser-Mercer B, Kunzli A, Korac M (1998) Prolonged turns in interpreting: effects on quality, physiological and psychological stress (Pilot Study). Interpreting: Int J Res Prac Interpreting 3:47–64

Normandin Y (1991) Hidden Markov models, maximum mutual information estimation and the speech recognition problem. PhD thesis, McGill University, Montreal, Quebec, Canada

Och FJ (2003) Minimum error rate training in statistical machine translation. In: 41st annual meeting of the Association for Computational Linguistics, Sapporo, Japan, pp 160–167

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29:19–51

Olszewski D, Prasetyo F, Linhard K (2005) Steerable highly directional audio beam louspeaker. In: Interspeech'2005 – Eurospeech, Lisboa, Portugal, pp 137–140

Papineni K, Roukos S, Ward T, Zhu W (2002) Bleu: a method for automatic evaluation of machine translation. In: 40th annual meeting of the Association of Computational Linguistics, Philadelphia, Pennsylvania, pp 311–318

Paulik M, Waibel A (2008) Extracting clues from human interpreter speech for spoken language translation. In: ICASSP 2008 IEEE international conference on acoustics, speech, and signal processing, Las Vegas, Nevada, pp 5097–5100

Ramabhadran B, Huang J, Picheny M (2003) Towards automatic transcription of large spoken archives – English ASR for the Malach project. In: Proceedings of the 2003 IEEE conference on acoustics, speech, and signal processing (ICASSP 2003), Hong Kong, China, pp 216–219

Ramabhadran B, Siohan O, Mangu L, Zweig G, Westphal M, Schulz H, Soneiro A (2006) The IBM 2006 speech transcription system for European Parliamentary speeches. In: Interspeech 2006 – ICSLP, ninth international conference on spoken language processing, Pittsburgh, PA, pp 1225–1228

Rao S, Lane I, Schultz T (2007) Optimizing sentence segmentation for spoken language translation. In: Interspeech 2007, 8th annual conference of the International Speech Communication Association, Antwerp, Belgium, pp 2845–2848

Renals S, Bengio S, Fiskus J (eds) (2006) Machine learning for multimodal interaction: third international workshop, MLMI 2006, Bethesda. Revised selected papers, LNCS 4299, Springer Verlag, Berlin

Rogina I, Schaaf T (2002) Lecture and presentation tracking in an intelligent meeting room. In: 4th IEEE international conference on multimodal interfaces (ICMI 2002), Pittsburgh, PA, pp 47–52

Roukos S, Graff D, Melamed D (1995) Hansard French/English, catalog nbr LDC95T20, Linguistic Data Consortium, Philadelphia, PA

Seleskovitch D (1978) Interpreting for international conferences: problems of language and communication. Pen & Booth, Washington DC

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of Association for Machine Translation in the Americas: visions for the future of machine translation, Cambridge, Massachusetts, pp 223–231

Soltau H, Metze F, Fügen C, Waibel A (2001) A one-pass decoder based on polymorphic linguistic context assignment. In: ASRU 2001, automatic speech recognition and understanding workshop, Madonna di Campiglio, Trento, Italy, pp 214–217

Soltau H, Yu H, Metze F, Fügen C, Jin Q, Jou S-C (2004) The 2003 ISL rich transcription system for conversational telephony speech. In: ICASSP 2004, IEEE international conference on acoustics, speech, and signal processing, Montreal, Canada, pp 773–776

Stolcke A (2002) SRILM – an extensible language modeling toolkit. In: 7th international conference on spoken language processing (ICSLP 2002, Interspeech 2002), Denver, Colorado, pp 901–904

Stüker S, Fügen C, Hsiao R, Ikbal S, Jin Q, Kraft F, Paulik M, Raab M, Tam Y-C, Wölfel M (2006) The ISL TC-STAR spring 2006 ASR evaluation systems. In: TC-STAR workshop on speech to speech translation, Barcelona, Spain, pp 139–144

Stüker S, Paulik M, Kolss M, Fügen C, Waibel A (2007) Speech translation enhanced ASR for European Parliament speeches – on the influence of ASR performance on speech translation. In: ICASSP 2007, international conference on acoustics, speech, and signal processing, Honolulu, Hawaii, pp 1293–1296

Trancoso I, Nunes R, Neves L (2006) Classroom lecture recognition. In: Vieira R, Quaresma P, Nunes MdGV, Mamede NJ, Oliveira C, Dias MC (eds) Computational processing of the Portuguese language, 7th international workshop, PROPOR 2006, Itatiaia, Brazil, LNCS 3960, Springer Verlag, Berlin, pp 190–199

Vidal M (1997) New study on fatigue confirms need for working in teams. Proteus Newsl NAJIT 6.1

Vogel S (2003) SMT decoder dissected: word reordering. In: International conference on natural language processing and knowledge engineering, Beijing, China, pp 561–566

Vogel S (2005) PESA: phrase pair extraction as sentence splitting. In: MT summit X, the tenth machine translation summit, Phuket, Thailand, pp 251–258

Vogel S, Ney H, Tillmann C (1996) HMM-based word alignment in statistical translation. In: COLING-96, the 16th international conference on computational linguistics, Copenhagen, Denmark, pp 836–841

Waibel A, Fügen C (2008) Spoken language translation. IEEE Signal Proc Mag 25(3):70–79

Waibel A, Stiefelhagen R (eds) (2009) Computers in the human interaction loop. Springer Verlag, Berlin

Waibel A, Jain AN, McNair AE, Saito H, Hauptmann AG, Tebelskis J (1991) JANUS, a speech-to-speech translation using connectionist and symbolic processing strategies. In: ICASSP-91, proceedings of the international conference on acoustics, speech, and signal processing, Toronto, Canada, pp 793–796

Waibel A, Steusloff H, Stiefelhagen R, the CHIL Project Consortium (2004) CHIL – computers in the human interaction loop. In: WIAMIS 2004, 5th international workshop on image analysis for multimedia interactive services, Lisbon, Portugal, 4 pp

Yagi SM (2000) Studying style in simultaneous interpretation. Meta J Traduc 45:520–547

Yuan J, Liberman M, Cieri C (2006) Towards an integrated understanding of speaking rate in conversation. In: Interspeech 2006 – ICSLP, ninth international conference on spoken language processing, Pittsburgh, Pennsylvania, paper Mon3A3O-1

Zechner K (2002) Summarization of spoken language – challenges, methods, and prospects. Speech Technol Expert eZine, 6

Zhan P, Westphal M (1997) Speaker normalization based on frequency warping. In: 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP'97), Munich, Germany, p 1039