Learning Speech Translation from Interpretation

zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
von der Fakultät für Informatik
Karlsruher Institut für Technologie (KIT)
genehmigte

Dissertation

von
Matthias Paulik
aus Karlsruhe

Tag der mündlichen Prüfung: 21. Mai 2010

Erster Gutachter: Prof. Dr. Alexander Waibel
Zweiter Gutachter: Prof. Dr. Tanja Schultz

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe sowie dass ich die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung des KIT, ehem. Universität Karlsruhe (TH), zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet habe.

Karlsruhe, den 21. Mai 2010

Matthias Paulik

Abstract

The **basic objective** of this thesis is to examine the extent to which automatic speech translation can benefit from an often available but ignored resource, namely human interpreter speech. The **main contribution** of this thesis is a novel approach to speech translation development, which makes use of that resource.

The performance of the statistical models employed in modern speech translation systems depends heavily on the availability of vast amounts of training data. State-of-the-art systems are typically trained on: (1) hundreds, sometimes thousands of hours of manually transcribed speech audio; (2) bi-lingual, sentence-aligned text corpora of manual translations, often comprising tens of millions of words; and (3) monolingual text corpora, often comprising hundreds of millions of words. The acquisition of such enormous data resources is highly time-consuming and expensive, rendering the development of deployable speech translation systems prohibitive to all but a handful of economically or politically viable languages. As a consequence, speech translation development for a new language pair or domain is typically triggered by global events, e.g. disaster relief operations, that incur a major need for cross-lingual, verbal communication—justifying the high development costs. In such situations, where an urgent need for cross-lingual communication exists, but no automatic speech translation solutions are (yet) available, communication is achieved with the help of human interpreters.

In this thesis, we introduce methods that exploit audio recordings of interpreter-mediated communication scenarios for speech translation system development. By employing unsupervised and lightly supervised training techniques, the introduced methods allow to omit most of the manual transcription effort and all of the manual translation effort that has typically characterized speech translation system development. Thus, we are able to significantly reduce the amount of time-consuming and costly human supervision that is attached to speech translation system development.

Further contributions of this thesis include: (a) a lightly supervised acoustic model training scheme for recordings of European Parliament Plenary Sessions, supporting the development of ASR systems in the various languages of the European Union without the need of costly verbatim transcriptions; and (b) a sentence segmentation and punctuation recovery scheme for speech translation, addressing the mismatch between output of automatic speech recognition and machine translation training data.

Zusammenfassung

Die vorliegende Dissertation[1] behandelt die Frage ob automatische Sprachübersetzung Nutzen aus Audioaufnahmen menschlicher Interpretationsszenarien ziehen kann. Im Kern der Arbeit werden Ansätze entwickelt, die es erlauben, die an der Sprachübersetzung beteiligten Komponenten, automatische Spracherkennung und maschinelle Übersetzung, mit Hilfe solcher Audioaufnahmen zu trainieren. Diese Ansätze werden anhand eines realen Anwendungsszenarios entwickelt, welches menschliche Simultanübersetzung (Interpretation), manuelle Transkription und manuelle Übersetzung im großen Stil verlangt: Sitzungen des Europaparlaments und die mit diesen Sitzungen verbundenen, multi-lingualen Dokumente. Die entwickelten Ansätze erlauben es, Sprachübersetzung direkt auf Aufnahmen menschlicher Interpretationsszenarien zu trainieren und benötigen dabei nur geringe Mengen an zeitaufwendiger und kostspieliger menschlicher Überwachung. Insbesondere wird nur ein geringer Teil der bisher für Sprachübersetzung notwendingen manuell transkribierten Sprachaufnahmen benötigt und keine der ansonsten notwendigen manuell angefertigten Übersetzungen.

Des weiteren wird im Rahmen dieser Dissertation ein Verfahren eingeführt, welches das Trainieren von Spracherkennungssystemen in den verschiedenen Sprachen der Europäischen Union unterstützt. Hierbei werden die frei zugänglichen Text- und Audioressourcen des Europaparlaments ausgenutzt, um akustische Modelle ohne kostspielige, wortgetreue Transkriptionen zu trainieren. Die vorliegenede Arbeit untersucht des weiteren, wie die Kombination von Spracherkennung und maschineller Übersetzung mit Hilfe einer automatischen Satzsegmentierung und einer automatischen Wiederherstellung von Satzzeichen verbessert werden kann.

[1] Appendix A beinhaltet eine Kurzfassung der Dissertation in deutscher Sprache.

Acknowledgements

First of all, I would like to thank my advisor Alex Waibel, whose vision inspired this thesis and who gave me the opportunity to conduct most of my research at the U.S. America part of interACT research, at Carnegie Mellon University. I greatly enjoyed my time in Pittsburgh, not the least because of the wonderful colleagues and the very friendly, warm and open atmosphere in our lab here at 407 S. Craig Street. Special thanks also go to my co-advisor Tanja Schultz for her valuable input, as well as to all the other people that had a direct impact on my thesis. This includes especially all the people that contributed to the TC-STAR SMT system build (Muntsin Kolss, Jan Niehues, Kay Rottmann, Stephan Vogel), the people that contributed with valuable technical discussions (Matthias Eck, Ian Lane, Kornel Laskowski, Thomas Schaaf), transcriptions (Susanne Burger and her team) and system components, including English acoustic models (Sebastian Stüker), German acoustic models (Christian Fügen and Matthias Wölfel) and a German compound word splitter (Florian Kraft). In fact, thinking back, the list of people I have to thank just goes on and on. Thanks simply have to go to all people at interACT research, USA and Germany alike!

Last, but certainly not least, I want to thank my parents and my wife, for supporting me, especially during those last stressful months.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

1.1 Motivation

Globalization as well as international crises and disasters spur the need for cross-lingual verbal communication for myriad languages. This is reflected in ongoing intense research activity in the field of automatic speech translation (ST). The field has seen tremendous performance improvements over the past two decades. Early efforts in ST started from the rather artificial research problem of translating speech recorded under controlled conditions, with restricted vocabularies, strong domain limitations and the necessity of a constrained speaking style. Nowadays, research in ST turned towards the task of translating spoken language as found in real life (spoken language translation) and constitutes as such one of the major research areas of speech and language processing. For example, major research projects of recent years focused on spoken language translation for the relatively broad domains of broadcast news and parliamentary speeches. The impressive advances in ST to date can largely be attributed to the statistical modeling schemes employed in the two component technologies of speech translation: automatic speech recognition (ASR) and machine translation (MT). Statistical modeling schemes for ASR and MT, and consequently ST, are primarily language independent and have been proven to work well for many language pairs. However, the performance of statistical models depends heavily on the availability of vast amounts of training data. Modern, large scale ST systems are typically trained on: (1) hundreds of hours of manually transcribed speech

audio; (2) sentence-aligned parallel text corpora, comprising tens of millions of manually translated words; and (3) monolingual text corpora, often comprising hundreds of millions of words. The high cost attached to acquiring such large amounts of training data turn out to be prohibitive for most language pairs and domains, limiting the availability of large-scale data collections to only a handful of languages. Consequently, ST development for a new language pair typically faces the problem of having no or only very limited training data resources readily available. As the resulting necessary data collection effort is not only cost intensive, but also highly time-consuming, deployable ST systems can only be made available for a new language after months or even years of effort. Such a delay is unacceptable for many situations that call for rapid development of automatic ST solutions, as given by disaster relief operations or military operations. The urgent need for cross-lingual, verbal communication in these situations, combined with the absence of automatic ST solutions, consequently necessitates the deployment of human interpreters.

In this thesis, we examine whether speech translation and its component technologies can benefit from human interpreter speech as a novel, low-cost data resource for system development. We develop methods to directly train speech translation systems on audio recordings of interpreter-mediated communication. By employing unsupervised and lightly supervised training techniques, the proposed methods allow us to omit most of the manual transcription effort and all of the manual translation effort that has typically characterized speech translation system development. Thus, the amount of costly and time-consuming human supervision necessary for speech translation system development is substantially reduced. We develop our methods on a large-scale, real-world spoken language translation task, for which large amounts of training data *are* available. This enables us to examine our approach under different levels of resource availability. We then transfer our most important findings to the setting of actual resource limitation, highlighting the feasibility and importance of our approach to developing speech translation systems for new language pairs rapidly and in a cost-effective manner. Further, the thesis also examines the question of how to optimally

combine ASR and MT for speech translation, following a (fully) decoupled ST architecture, as described in Section 3.4.

1.2 Outline

The following chapter, Chapter 2, discusses the background and related work. All subsequent chapters will be presented in two parts.

Chapters 3 to 6 are a description of the basic overall experimental setup that is used for most of the experiments conducted within this thesis. These chapters introduce the basic methods for automatic speech translation applied throughout this work. Further, they describe the spoken language translation task on which most our experiments are based, and they also describe the ASR and MT systems that were developed for this task. In this context, we also describe our experiments to improve the *combination* of ASR and MT for automatic speech translation of spoken language.

The second part of this thesis is presented in Chapters 7 to 12. Here, we describe experiments that aim to exploit audio recordings of human interpreter-mediated communication as a novel resource for speech translation system development. Finally, in Chapter 13, we summarize and discuss our results, and we identify some of the remaining research challenges. A more detailed overview on the chapters of the two main parts of the thesis is given in the following.

1.2.1 Part I

Chapter 3, reviews the statistical methods applied in state-of-the-art speech translation and its component technologies, ASR and MT. The chapter also gives a short summary of some of the algorithms and implementations resulting from theoretical statistical formulations and it further describes the performance metrics used throughout this work. Chapter 4 introduces the large-scale spoken language translation task that will set the stage for most of our experiments: the speeches

of the European Parliament as well as the recordings of the simultaneous interpreters supporting the sessions of the European Parliament. In Chapter 5 we explain, with our English and Spanish ASR and MT systems as an example, the training and decoding schemes as they are used throughout this work.

Chapter 6 describes our experiments to improving the combination of ASR and MT for speech translation of spoken language. Specifically, we describe in this chapter our sentence segmentation and punctuation recovery scheme for spoken language translation.

1.2.2 Part II

In chapter 7, we take a first look at interpretation as a data resource for speech translation by closely examining the nature of interpretation and comparing it to manual and automatic translation. We identify several possible ways to exploit interpretation as a data resource for improving ST performance. One presented idea involves exploiting interpretation as an auxiliary information source, by biasing ASR and MT, applied to source language speech, with information extracted from already available interpretation in the target language. Experiments based on this idea are presented in Chapter 8.

Chapters 9, 10 and 11 examine interpretation audio as training data resource for ST. Specifically, Chapter 9 examines interpretation audio, as it is available for sessions of the European Parliament, for acoustic model training. Chapter 10 introduces our approach for training translation models from interpreter speech. In Chapter 11, we present a framework that allows for an automatic training data extraction from interpretation audio and a successful application of such extracted training data, by tying together the approaches developed in the previous chapters. We conclude our experiments in Chapter 12, by transferring our most important findings to a setting of actual resource limitation: speech translation development between English and Pashto.

2

Related Work

The main focus of this thesis lies on the development of human interpreter speech as a novel resource for building speech translation systems. Chapter 6, which introduces a sentence segmentation and punctuation recovery scheme for spoken language translation, deviates from this main focus. For this reason, the discussion of work related to sentence segmentation and punctuation recovery is presented within Chapter 6. In the following we shortly discuss work that is related to the main objective of this thesis.

We are not aware of any previous work on exploiting human interpreter speech for training automatic speech translation. This is particularly true for our exploitation of *interpretation* audio for training automatic *translation* (models). However, this work is related to and stems from ideas first presented in 1994 and 1995 by Brown et al. [9] and Brousseaux et al. [8], respectively. Both propose to improve dictation systems for professional translators with the help of information that is automatically derived from the source language text that is to be translated. This scenario has seen renewed interest in recent years [3; 30; 53]. While all these previous works only considered biasing dictation systems with knowledge extracted from source language *text*, we applied the described approach in [53; 54] for the first time to extract knowledge from source language *speech*. However, our experiments presented in [53; 54] only considered read-speech, with source and target language speaker reading from a travel-domain parallel text corpus of sentence-aligned translations. In contrast to this rather

artificial task, we consider in this thesis speech audio as it occurs in real-world human interpretation. Further, we do not aim to develop or improve a dictation system for human translators, but we aim to train speech translation systems using such interpretation speech audio.

2.1 Limiting the Amount of Human Supervision

The enormous training data requirements of the statistical methods governing the component technologies of speech translation, automatic speech recognition and machine translation, have prompted numerous research trying to limit the amount of costly human supervision attached to the creation of such training data. The in this thesis developed approaches for exploiting interpreter speech for speech translation aim to significantly limit the amount of costly human supervision necessary for ST development. This thesis therefore needs to be seen in the context of previous research that aims to limit the amount of supervision for ASR or MT.

2.1.1 Limiting Supervision in Automatic Speech Recognition

Unsupervised and lightly supervised acoustic model training [37] are common approaches in automatic speech recognition to limit the amount of costly human supervision. Unsupervised acoustic model training is based on large amounts of speech data for which no human transcriptions are available. Training relies on automatic transcriptions that are created with an initial ASR system that was trained on small amounts of manually transcribed speech audio. Lightly supervised acoustic model training refers to the case where some imperfect human transcriptions, for example closed-captions provided during television broadcasts, can be used to either bias the initial ASR system for an improved transcription performance or to filter erroneous ASR hypotheses.

The application of language independent and language adaptive acoustic models [66] is another possibility to limit the amount of manually transcribed audio data needed for training accurate acoustic models. The core idea here is to limit the necessary amount of transcribed speech data for a new language by borrowing models and data from one or more other languages.

Similar to exploiting automatic transcriptions for unsupervised acoustic model training, it is also possible to exploit automatic transcriptions as additional language model training data [45]. In situations were only limited amounts of in-domain text data are available for language modeling, it is also possible to automatically collect additional in-domain data from the world-wide-web [85]. However, this approach is only feasible for the handful of languages were large amounts of monolingual text data are available via the world-wide-web.

2.1.2 Limiting Supervision in Machine Translation

Similar to collecting monolingual text data from the world-wide-web for language modeling, it is possible to crawl the web for comparable corpora [19] that can be used for translation model training. Comparable corpora are bilingual texts that are not translation of each other, but that are related and include the same information to some degree. An example for comparable corpora are the online articles of news agencies in different languages. As in the case of crawling monolingual text data for language modeling, collecting comparable corpora is again limited to only the major languages of the world-wide-web. Further, as the major source for comparable corpora are the web pages of news agencies, the domain of comparable corpora is mostly limited to news.

The analogoue to unsupervised acoustic model training, namely unsupervised translation model training, was first investigated by Ueffing et al. in [76]. Ueffing et al. refer to this concept as 'self-training'. As only a very small amount of monolingual data were used for self-training, the approach was presented more in the context of domain adaptation, rather than unsupervised training. In detail, Ueffing et al. applied machine translation only to a test set, and then selected

the most reliable automatic translations to build a small phrase table. This small phrase table was then used together with the baseline phrase table to re-translate the test set in a second pass. Self-training (unsupervised translation model training) in the context of large monolingual corpora was first investigated in [67].

Another possibility to limit the amount of costly human supervision in the context of machine translation is to reduce the amount of necessary parallel text data as strong as possible, without impacting automatic translation performance. Eck et al. [15] sort the sentences of monolingual text data, with the top n sentences representing the most valuable sentences for translation model training. Only these top n sentence are then given to human translators to create parallel text.

PART I

State-of-the-art Speech Translation
&
a Large-Scale Spoken Language Translation Task: European Parliament Plenary Sessions

3

Statistical Speech Translation

3.1 Terminology

Automatic speech recognition (ASR) converts speech to text. *Machine translation* (MT) refers to the automatic translation of source language text to target language text. *Speech translation* (ST) refers to the automatic translation of source language speech to target language text, for example by applying machine translation to the output of automatic speech recognition. In the context of speech translation, two additional terms are frequently used; *speech-to-speech translation* (S2S) and *spoken language translation* (SLT). In speech-to-speech translation, the output modality is speech rather than text, achieved with the help of speech synthesis systems. Spoken language translation refers to speech translation that is applied to spoken language 'as found in real life', which often suffers from speech disfluencies like fillers, repetitions and corrections. Examples for 'real life' speech are parliamentary speeches or the conversational speech encountered in television shows.

3.2 Automatic Speech Recognition (ASR)

State-of-the-art ASR systems are based on statistical methods. The *fundamental equation of speech recognition* applies Bayes' decision rule to rewrite the classification problem of finding the most likely word sequence $\hat{S}$ given the observed

sequence X of feature vectors (extracted from the acoustic signal) as follows:

$$\begin{aligned}\hat{S} &= \arg\max_S P(S|X) \\ &= \arg\max_S \frac{P(S)P(X|S)}{P(X)} \\ &= \arg\max_S P(S)P(X|S)\end{aligned} \tag{3.1}$$

By applying Bayes' theorem, a decomposition into two independent probability distributions is achieved. The language model (LM) $P(S)$ determines the prior probability of observing the word sequence S, and the acoustic model (AM) $P(X|S)$ represents the probability of observing the sequence X of feature vectors given S. State-of-the-art ASR systems typically apply n-gram language models and Hidden Markov acoustic models [61]:

- N-gram language models provide the likelihood of the word w_i, given a history of words $w_1...w_{i-1}$, by approximating it with the likelihood of w_i given only the $n-1$ preceding words. In the case of a tri-gram LM, the probability of w_i is therefore given as: $P(w_i|w_1...w_{i-1}) = p(w_i|w_{i-2}, w_{i-1})$.

- Hidden Markov Models (HMMs) are stochastic finite state-automata, consisting of a Markov chain of states. Each hidden state has an emission probability distribution of observable output tokens. In the context of speech recognition, HMM states are acoustic states and the output tokens are the observable acoustic feature vectors. Most commonly, the state-specific emission probability distribution is modeled with the help of Gaussian Mixture Models. The transitions between the states, together with their transition probabilities, serve to model the temporal structure of speech [44; 61]. In ASR, HMMs are used to model sub-word units, typically phoneme units. Figure 3.1 gives an example for a three-state HMM modeling the phoneme /æ/. The model consists of three sub-phoneme acoustic states in a strictly sequential left-to-right topology. Word models are constructed, as they become needed, by concatenating phoneme models, as described in more detail in Section 3.2.2.

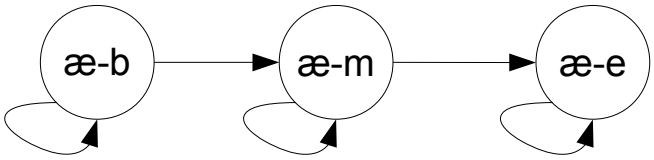

Figure 3.1: Example for a Hidden Markov Model phoneme unit.

3.2.1 Language Model Training

Language model training data consists of text corpora comprising often hundreds of millions of words. These text corpora are used to estimate the n-gram probabilities $p(w_i|w_{i-2}, ..., w_{i-n+1})$, based on their occurrence counts:

$$p(w_i|w_{i-n+1}, ..., w_{i-1}) = \frac{\#(w_{i-n+1}, ..., w_i)}{\#(w_{i-n+1}, ..., w_{i-1})} \tag{3.2}$$

For uni-grams, the probability is given by:

$$p(w_i) = \frac{\#w_i}{\sum_j \#w_j} \tag{3.3}$$

Even the largest training corpora do not contain all possible n-gram combinations that are valid for a specific ASR vocabulary. To avoid zero probabilities for such unseen n-grams, smoothing (also known as discounting) in combination with LM back-off to shorter word histories is applied. Discounting means that some probability mass from the observed n-grams is removed and redistributed to the unobserved n-grams. In the context of this work, we apply Kneser-Ney smoothing [34].

3.2.2 Acoustic Model Training

Acoustic model training data consists of large amounts, often hundreds of hours of speech audio, transcribed at the word level. To adjust the parameters of the HMMs so that the acoustic models 'optimally'[1] represent the sequences of feature

[1]Different optimization criteria are used for AM training, e.g. Maximum Likelihood or Maximum Mutual Information.

vectors found in the training data, it is first necessary to transform the training word sequences (the transcriptions) into sequences of HMM states. This is accomplished with the help of pronunciation dictionaries that list the phonetic transcription of words. Word models are build by concatenating the HMM phoneme units in a linear fashion, from left-to-right, to form word models, and ultimately word sequences. Given these word sequence HMMs together with their observed sequences of feature vectors it is now possible to optimize the HMM parameters. Most commonly, the Baum-Welch algorithm (in the context of Gaussian Mixture Models) is applied for parameter optimization. The Baum-Welch algorithm [4] is a special case of the Expectation Maximization [13] algorithm and applies the Maximum Likelihood (ML) optimization criterion. Modern ASR systems typically also apply discriminative training methods after ML training, as for example Maximum Mutual Information [58] training or Minimum Phone Error [65] training.

3.2.3 Decoding

The task of finding the word sequence $\hat{S}$ that maximizes equation 3.1 is accomplished during the so-called decoding stage. Decoding can be imagined as the task of finding the best possible path through a search graph, consisting of a huge HMM that represents all possible word sequences S. This search graph combines the acoustic model probabilities with the language model probabilities, by applying the LM at transitions between words. This means that the score (negative logarithm of probabilities) of each path through this search graph is computed by accumulating the individual (and usually differently weighted) AM and LM scores. Different decoding strategies are applied in modern ASR systems. In the following, we will shortly describe the decoding strategy applied by the IBIS single-pass decoder [71] (part of the Janus Recognition Toolkit, JRTk [16]) as we use this decoder throughout this work.

The IBIS decoder applies a time-synchronous Viterbi beam search for decoding. The search for $\hat{S}$ is conducted in a dynamically constructed search graph. To dynamically build this search graph, the decoder evaluates the presented

speech frames sequentially[1]. The Viterbi approximation (maximum approximation) helps to limit the computational overhead. In each time step, the search states in the graph are updated only by the score of the best incoming path, instead of considering all incoming paths. To deal with the combinatorial explosion associated with large vocabularies, typically only the best states (hypotheses) are expanded in each time step—this heuristic is commonly referred to as 'beam search'.

3.3 Statistical Machine Translation (MT)

Statistical machine translation is based on the same basic statistical methods as ASR. Brown et al. [10] introduce the *fundamental equation of statistical machine translation* as:

$$\begin{aligned} \hat{T} &= \arg\max_T P(T|S) \\ &= \arg\max_T \frac{P(T)P(S|T)}{P(S)} \\ &= \arg\max_T P(T)P(S|T) \end{aligned} \tag{3.4}$$

The most likely word sequence $\hat{T}$ in the target language given a word sequence S in the source language can be computed with the help of two independent models: the target language model $P(T)$ and the translation model $P(S|T)$. As in ASR, MT typically applies n-gram language models. Virtually all statistical translation models use the IBM alignment-lexicon models [10] as a starting point [11]. These models provide the translation probability $p(t|s)$ of the source and target word pair (t, s). The probabilities are estimated on large amounts of sentence-aligned, bilingual parallel text of manual translations—often comprising tens of millions of translated words. Similar to the hidden states of HMMs in ASR, the concept of *word alignment* is used to describe the unknown correspondences between source and target words [11] (which word in the training source

[1]The original audio signal is represented by a sequence of speech frames. The typical frame size is 10ms.

sentence S corresponds to which word in the respective training target sentence T).

A generalization of Equation 3.4 can be achieved by directly modeling the posterior probability $P(T|S)$ in a log-linear framework, as proposed by [47; 51]. Here, $P(T|S)$ is given by different models $M_i(T|S)$ and their scaling factors λ_i as follows:

$$P(T|S) = \frac{exp(\sum_i \lambda_i M_i(T|S))}{\sum_{T'} exp(\sum_i \lambda_i M_i(T'|S))} \tag{3.5}$$

The denominator only depends on the source sentence S. Therefore, $\hat{T}$ can be expressed as:

$$\begin{aligned} \hat{T} &= \arg\max_T P(T|S) \\ &= \arg\max_T exp(\sum_i \lambda_i M_i(T|S)) \end{aligned} \tag{3.6}$$

This generalized approach allows for an easy integration of additional models M_i. The scaling factors λ attached to these models are typically trained using minimum error rate training [46], as it is explained in more detail at the end of Section 5.2.

3.3.1 Phrase-Based Approach

Throughout this work, we rely on phrase-based statistical MT, which can easily be identified as today's most popular approach to statistical MT. Instead of translating a source sentence on a source-word to target-word basis into the target sentence, the basic idea of phrase-based MT is to translate source phrases, comprised of one or more words, into target phrases. During decoding, this involves segmenting the source sentence into source phrases and then composing the target sentence from the translated phrases. Phrase pairs are typically extracted from the training data based on IBM model word alignments that are computed for both translation directions. Various phrase extraction methods are used today. Throughout this work, we rely on the phrase extraction method described by

Koehn et al. [35]. The extracted phrases are stored in so-called phrase-tables. Each line in such a phrase-table file includes one phrase-pair accompanied with the phrase-pair probabilities, as determined by the applied translation models M_i.

The search for the target sentence $\hat{T}$ during decoding applies the important constraint that '*all* positions of the source sentence should be covered exactly *once*' [11]. Several operations have to be taken into account during decoding: segmenting the source sentence into phrases, reordering the phrases in the target language, and determining the most probable word sequence [11]. Various decoding strategies are available. In the following, we describe the search strategy of the Interactive Systems Laboratories (ISL) beam search decoder [78], since we use this decoder throughout this work:

Here, decoding is organized in two stages. First, a so-called translation lattice is build, after which a best path search is conducted through the lattice. Starting from the source sentence, a linear graph is constructed that includes one edge per source word. Then, additional edges are inserted into the graph, according to the phrase translations presented in the phrase table. To limit the search space, only the n best translation alternatives are considered during phrase insertion. The best path search during the second step includes the application of additional models not included in the phrase table, as for example the language model. Further, an internal word reordering model is applied that allows for word permutation within a limited reordering window. The search space is pruned by applying a relative beam.

3.4 Speech Translation: Combining ASR and MT

Speech Translation can be viewed as the problem of combining its component technologies, ASR and MT, in a computational feasible way for an optimal translation performance. Throughout this work, we follow the popular decoupled

approach to statistical speech translation, which relies on a *sequential approximation* of the joint optimization problem. In the following, we give a formal motivation to the sequential approximation, based on the formulations presented in [11].

Formally, we seek the target sentence $\hat{T}$ given a sequence of source language acoustic feature vectors X. By introducing the source sentence S as a hidden variable and by assuming that $P(T|S,X) = P(T|S)$ we can write:

$$\begin{aligned}\hat{T} &= \arg\max_T P(T|X) \\ &= \arg\max_T \sum_S P(T,S|X) \\ &= \arg\max_T \sum_S P(T|S)P(S|X) \\ &\cong \arg\max_T \{\max_S P(T|S)P(S|X)\}\end{aligned} \tag{3.7}$$

In the last step, we applied the maximum approximation. The two tasks involved in speech translation, ASR and MT, are clearly represented in this formulation by the posterior probabilities $P(S|X)$ and $P(T|S)$. We can now fully decompose the speech translation task by applying the following sequential approximation:

$$\begin{aligned}\hat{T} &= \arg\max_T \{\max_S P(T|S)P(S|X)\} \\ &\cong \arg\max_T P(T|\arg\max_S P(S|X))\end{aligned} \tag{3.8}$$

Figure 3.2 depicts the speech translation system setup that follows from this sequential approximation. The drawback of the sequential approximation is obvious: MT is simply applied on the error-prone first-best ASR hypothesis, resulting in an accumulation of ASR and MT errors. For this reason, numerous works have investigated to enrich the interface between ASR and MT with the competing n-best ASR hypotheses, which may contain more accurate results. For example it was investigated to apply MT to ASR n-best lists [83], ASR word-lattices [41] or

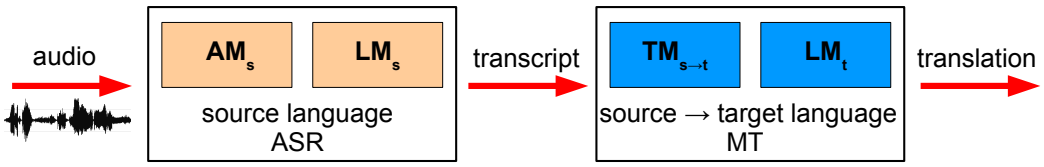

Figure 3.2: Speech translation system setup following the decoupled approach.

ASR confusion networks [5]. Another problem results from a mismatched condition between typical MT training data and ASR input to the translation system. Most available MT training data consist of well-formed sentences with proper punctuation. ASR output, on the other hand, does not include punctuation, suffers from recognition errors and is usually segmented using voice activity detection. Further, speech translation that is applied to spoken language as encountered in real life (spoken language translation), suffers from ill-formed sentence structures and frequent speech disfluencies (fillers, repetitions and corrections). To tackle this mismatch, it is possible to enrich the interface between ASR and MT with an ASR post-processing component that removes fillers, re-segments ASR output into more sentence-like units and introduces punctuation.

For further reading: an excellent overview on the state-of-the-art in automatic speech translation is given by articles presented in 'IEEE Signal Processing Magazine: Special Issue on Spoken Language Technology', May 2008 [11; 20; 80].

3.5 Performance Evaluation

The primary performance metrics used in this work are word error rate (WER) for measuring ASR performance and BLEU metric [50] for measuring MT and ST performance.

Word error rate is based on the minimal edit distance between hypothesis and reference sentence, this means it is based on the minimal number of substitutions s, insertions i, and deletions d necessary to transform the hypothesis into the reference. With n the number of reference words, the WER is given as:

$$WER = \frac{s + i + d}{n} \times 100\% \tag{3.9}$$

BLEU metric [50] compares the MT hypotheses to one or more human reference translations based on n-gram comparison. Specifically, it computes the geometrical mean of the modified n-gram precisions with $n \in \{1, ..., 4\}$ and applies a length penalty to translation hypotheses that are shorter than the, in regard to its length, best matching reference translation. The n-gram precisions are modified in a way to serve the intuitive demand for considering a reference n-gram as exhausted after a matching candidate n-gram is identified. In its original definition, the BLEU score ranges from 0 to 1, whereas a translation that is identical to a reference translation attains a score of 1. However, throughout this work the BLEU score will be given in the range from 0–100, i.e. multiplied by the factor 100. Depending on the used BLEU scoring script, we identify two different 'versions' of BLEU metrics in this work. Whenever we use the BLEU scoring script 'mteval-v11b', provided by NIST, we speak of NIST BLEU metric, otherwise of IBM BLEU metric. NIST BLEU incorrectly computes the brevity penalty based on the length of the shortest reference translation, while IBM BLEU computes the brevity penalty based on the closest matching reference translation.

4

A Large-Scale Spoken Language Translation Task: European Parliament Plenary Sessions (EPPS)

4.1 Task Description

The European Union (EU) language policy actively promotes the freedom of its citizens to speak and write their own language. This is reflected by the fact that "legislation and documents of major public importance or interest are produced in all 23 official languages" of the Union and that all other documents are translated into "the languages needed" [77]. This multilingualism entails a tremendous translation and interpretation effort, costing the EU more than 1 billion Euros annually [77]. A good example for this tremendous effort is the European Parliament. The document flow of the European Parliament is fully multilingual from the outset, as Members of the Parliament need working documents in their own language [77]. This means that proceedings of European Parliament Plenary Sessions (EPPS) are translated into all 23 official languages. These so-called final text editions (FTEs) are made publicly available approximately 2 months [24] after a session on the Parliamentary web pages. Further, it is necessary to provide Members of the Parliament with a simultaneous interpre-

tation in their native tongue, as speeches in the Parliament can be given in any of the official languages. These interpretations, along with the original politician speech, are broadcast live via satellite on different audio channels. Figure 4.1 depicts the described translation and interpretation effort. The live broadcast, language dedicated audio channels are shown on the upper right hand side of Figure 4.1. An interpreter provides the simultaneous interpretation in language L_i whenever a politician is speaking in a language $L_{j \neq i}$. In the case that a politician is speaking in the respective language of an audio channel, the original speech of the politician is being broadcast on that channel.

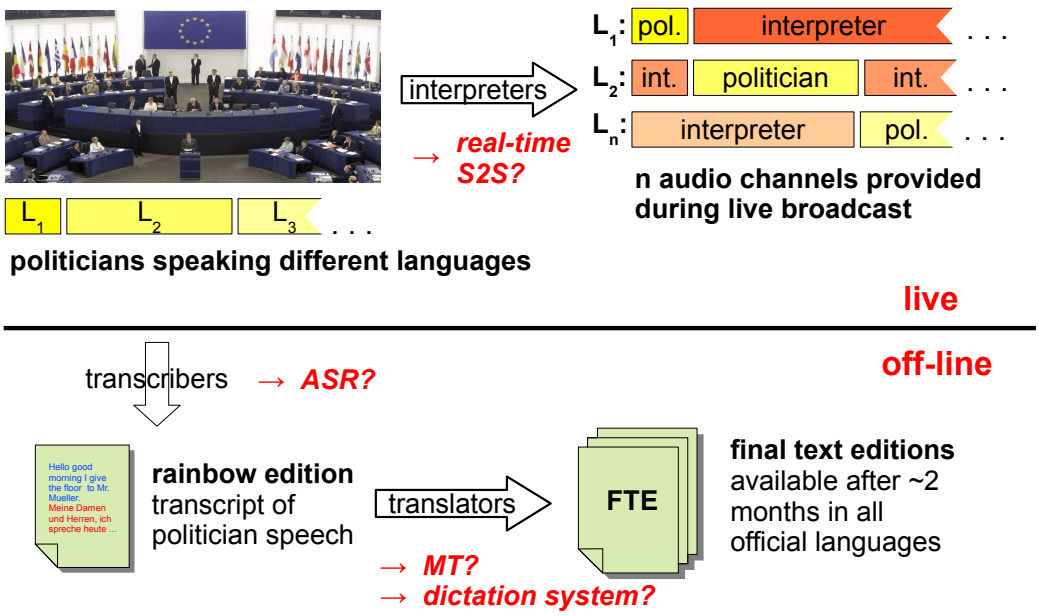

Figure 4.1: Manual transcription, translation and interpretation in the context of European Parliament Plenary Sessions (EPPS) and possible automatic solutions: automatic speech recognition (ASR), machine translation (MT) and speech-to-speech translation (S2S).

As shown in Figure 4.1, the tremendous translation and interpretation effort necessary for EPPS offers various possible application scenarios for speech translation and its component technologies, ASR and MT. For example, real-

time speech-to-speech translation (S2S) could be employed during EPPS, easing the interpretation effort. Further, it is possible to support the creation process of the final text editions by a) applying ASR for an automatic transcription of the speeches held in the Parliament and/or b) by automatically translating these transcriptions, or their revisions[1], into the various languages of the Union. In addition to MT, automatic speech recognition can further support the translation process in form of dictation systems for human translators. Such dictation systems can significantly speed up the manual translation process by allowing the human translator to dictate the translation, rather than typing it. Another application scenario is the automatic translation of manually or semi-automatically created FTEs. For example, one manually created English FTE could be automatically translated into all other languages of the Union.

In the context of this work, we mainly concentrate on the task of automatically transcribing and translating the speeches given by the politicians[2]. Such **verbatim transcriptions and translations** are valuable for archiving purposes and they can also directly support the creation process of the FTEs. We develop various scientific approaches for this verbatim transcription and translation task. The developed approaches are not just valuable for spoken language translation in the context of EPPS, but more importantly, they are valuable for automatic speech translation in general.

4.2 EPPS Data Resources

Significant data resources are available in the context of EPPS. The final text editions published on the web pages of the European Parliament in the various languages of the Union are an ideal data resource for training statistical language models and translation models. For example, the publicly available Europarl corpus [32] combines FTEs in 11 European languages in a multilingual, sentence-aligned text corpus for the development of statistical machine translation systems.

[1] Politicians are allowed to revise the transcriptions of their speeches held in Parliament.

[2] For this task, it is also beneficial to automatically transcribe and translate the EPPS simultaneous interpretations, as explained in the following chapters.

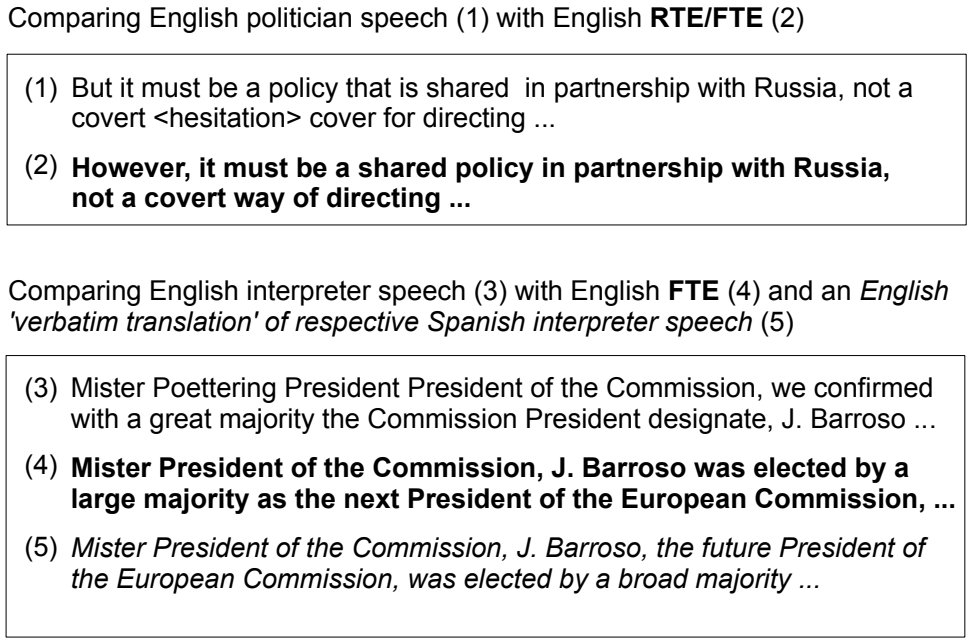

Figure 4.2: Comparing rainbow text edition (RTE) and final text edition (FTE) with respective verbatim transcription and translation of politician and interpreter speech.

In its current version, the Europarl corpus v3 includes FTEs from 1996 to 2006 and comprises up to 55 million words per language [31]. Further, the live broadcast audio channels, as they are depicted in Figure 4.1, can be recorded and used for acoustic model training. Within the project TC-STAR (compare Section 4.3), University RWTH Aachen recorded approximately 100h of English and Spanish live broadcast EPPS audio for the purpose of ASR development. Despite the existence of the so-called rainbow editions (RTEs), depicted in Figure 4.1, and the final text editions, verbatim transcriptions had to be created for the recorded English and Spanish speech[1]. As explained in Section 4.1, the language dedicated audio channels contain a mix of politician speech and interpreter speech. The RTEs only provide a transcription of politician speech, "aim for high readability" [24] and include revisions made by the politicians themselves. Therefore, they do not provide a verbatim transcript but differ, in parts strongly, from the original

[1]RWTH Aachen and UPC provided these verbatim transcriptions within TC-STAR.

politician speeches. Gollan et al. [24] note that the RTEs include the removal of hesitations, false starts and word interruptions. Furthermore, they note that transpositions, substitutions, deletions and insertions of words can be observed. The final text editions, which are solely based on the rainbow editions, consequently include these differences between original politician speech and RTE. Furthermore, the translations found in the FTEs differ even more strongly from verbatim transcriptions of respective interpreter speech. Figure 4.2 gives an example in which we compare (a) a verbatim transcript of English politician speech with its RTE/FTE transcript and (b) a verbatim transcript of English interpreter speech with its respective section in the FTE. For the latter example, we also show a manual English 'verbatim translation' of Spanish interpreter speech. In addition to the described data resources, we have in-house collected recordings of live broadcast EPPS available. Our recordings include several of the broadcast language dedicated audio channels, including German, English and Spanish.

4.3 The European project TC-STAR

The European project "Technology and Corpora for Speech-to-Speech Translation" (TC-STAR) was a three year project that aimed to advance research in the core technologies for speech-to-speech translation. Three competitive evaluations were conducted in the years 2005, 2006 and 2007 to foster advances in all speech-to-speech translation technologies. The evaluations were split in four sub-categories. These sub-categories included evaluations in (a) automatic speech recognition, (b) machine translation, (c) text-to-speech and (d) end-to-end performance. TC-STAR participating sides competed in the sub-categories (a) to (c). End-to-end performance was not measured on individual systems or on a system build from the best competing ASR and MT systems. Instead, system combination techniques where applied whenever possible. For example, the output of the ASR systems that had been competing in category (a) was combined by applying NIST's Recognizer Output Voting Error Reduction (ROVER) [17] before handing it over to MT for spoken language translation.

4.3.1 EPPS Machine Translation tasks of TC-STAR

TC-STAR distinguished between three machine translations tasks in the context of EPPS. These tasks considered machine translation between English and Spanish. The **SLT task** constituted the task of automatically translating the ROVERed English and Spanish ASR system outputs. In the **verbatim task**, the respective manual verbatim transcriptions replaced the ROVERed ASR hypotheses as input to the MT systems. Finally, the **FTE task** considered the automatic translation of English and Spanish final text editions.

4.4 EPPS Development and Evaluation Sets

For performance evaluation of our EPPS English and Spanish ASR and MT systems, we rely on the development and evaluation sets as they were provided within the project TC-STAR. Specifically, we make use of the 2005 development set, the 2006 development set (dev06) and the 2007 evaluation set (eval07). These sets include case-sensitive transcription and translation references with proper punctuation. For measuring translation performance, two reference translations are provided. Our default scoring of ASR and MT performance relies on case-insensitive WER and case-insensitive BLEU, respectively. Further, we usually remove the punctuation marks present in the transcription and translation references before scoring[1]. We explicitly state whenever we apply a case-sensitive scoring or a scoring that includes punctuation.

Dev06 and eval07 only comprise politician speech and no interpreter speech. Further, no manual speech utterance segmentation is provided for dev06 and eval07. In the context of this work, we use a language-independent HMM based speech/non-speech audio segmentation to infer speech utterances before applying ASR. For unsupervised speaker-adaptation, we cluster the resulting speech utterances into several speaker clusters, using the hierarchical, agglomerative clustering technique described in [29]. In contrast to dev06 and eval07, the TC-STAR

[1] All data statistics listed in this section refer to non-punctuated references.

2005 development set presents a mix of politician and interpreter speech. Further, it is provided with a manual utterance segmentation that is kept consistent for the reference transcriptions and the reference translations. In other words, for all English and Spanish speech utterances included in the 2005 set, aligned transcription and translation references are provided. As these properties of the TC-STAR 2005 development set are of special value for the experiments described in Chapter 8 and 10, we extract two European Parliament sessions from it, for further development and evaluation purposes. In the following, we refer to these two sets simply as dev05 and dtest05. Table 4.1 lists the data statistics of the English and Spanish dev05 and dtest05 sets. Table 4.2 shows the data statistics for dev06 and eval07. Due to the automatic utterance segmentation applied prior to ASR on dev06 and eval07, the amount of speech segments and reference translations segments differs. To align the translated speech utterances to the translation reference for scoring SLT performance, we rely on the multiple reference word error (mWER) [42] script, as it was provided by RWTH Aachen within TC-STAR.

	English		**Spanish**	
	dev05	dtest05	dev05	dtest05
utterances	1256	448	1589	849
words [k]	17.4	5.9	14.7	6.6
audio [min]	95	40	89	40

Table 4.1: Data statistics for dev05 and dtest05.

	English		**Spanish**	
	dev06	eval07	dev06	eval07
speech utt.	1287	1926	1707	2085
transl utt.	1194	1167	792	746
words [k]	27.9	26.0	22.4	25.8
audio [h]	3.2	2.7	2.3	2.7

Table 4.2: Data Statistics for dev06 and eval07.

4. A LARGE-SCALE SPOKEN LANGUAGE TRANSLATION TASK: EUROPEAN PARLIAMENT PLENARY SESSIONS (EPPS)

5

EPPS English/Spanish Automatic Speech Recognition and Machine Translation

In the following sections, we explain on the example of our English/Spanish systems the ASR and MT training and decoding schemes as they are used throughout this work.

5.1 Automatic Speech Recognition Systems

Our ASR systems are developed with the Janus Recognition Toolkit (JRTk), featuring the IBIS single pass decoder [71]. For language model estimation, we rely on the SRI Language Model toolkit [72].

5.1.1 Front-ends

Our preprocessing typically relies on traditional Mel-frequency Cepstral Coefficients (MFCC). In some cases, we also apply a pre-processing that is based on the warped minimum variance distortionless repsonse (MVDR). The latter replaces the Fourier transformation by a warped MVDR spectral envelope [81] . Our front-ends provide features every 10ms. However, for speaker adaptive decoding in a multi-pass setup, we change this value to 8ms. The applied front-ends use

13 cepstral coefficients, with mean and variance normalized on a per utterance basis. Seven adjacent frames are combined into one single feature vector which is then reduced to 42 dimensions using linear discriminant analysis.

5.1.2 Acoustic Model Training

Our ASR systems are based on sub-phonetically tied three-state Hidden Markov Models without state-skipping. The applied acoustic model training scheme first estimates context independent AMs and then uses JRTk's standard top-down clustering approach to obtain context-dependent models. Training involves incremental splitting of Gaussians and several iterations of Viterbi training. For the English ASR system in Chapter 12, we also apply boosted Maximum Mutual Information training [59]. The systems feature single global semi-tied covariance matrices after linear discriminant analysis [23]. In the case of unsupervised speaker adaptation in a second decoding pass, we employ acoustic models trained via feature space speaker adaptive training.

5.1.3 English Automatic Speech Recognition

The EPPS English ASR system is based on ASR sub-systems that were developed at University Karlsruhe by Stüker et al. [74] for the TC-STAR 2006 evaluation. The featured decoding setup is a simplified version of the 2006 evaluation system's decoding setup.

The decoding setup consists of two decoding passes, with two ASR sub-systems in each pass and confusion network combination [40] at the end of each pass. One MFCC front-end based ASR sub-system and one MVDR front-end based ASR sub-system is used in each decoding pass. The ASR sub-systems of the first pass apply speaker-independent AMs while the sub-systems of the second pass apply feature space speaker adaptive training models. Unsupervised speaker adaptation is performed on the output of the first decoding pass by applying Maximum Likelihood Linear Regression (MLLR) [38], feature space constrained MLLR [22] and Vocal Tract Length Normalization [82]. Figure 5.1 depicts the decoding setup and list the eval07 word error rates achieved in the

several decoding passes. The acoustic models of the ASR sub-systems are trained on approximately 80h of English EPPS data, as provided by RWTH Aachen for the TC-STAR 2006 evaluation. The pronunciation dictionary consists of 47k lowercased pronunciation entries. The 4-gram LM is trained on the 2006 available EPPS transcriptions (750k words) and EPPS final text editions (33M words) as well as on the Hub4 broadcast news data (130M words) and the English part of the UN Parallel Text Corpus v1.0 (41M words). Table 5.1 lists the language model perplexity (PPL), out-of-vocabulary (OOV) rate and case-insensitive WER on our dev and eval set (please refer to Section 4.4 for a description of dev and eval).

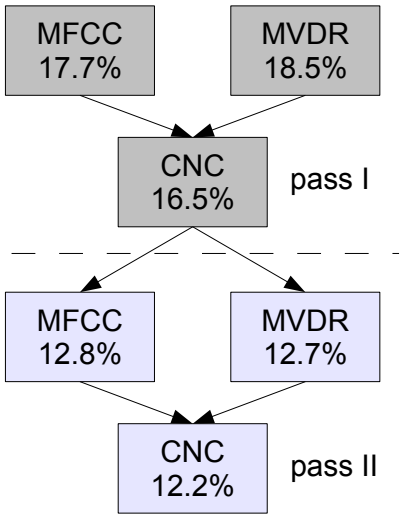

Figure 5.1: ASR decoding setup and influence on word error rate (English, eval07). The setup applies two decoding passes with ASR systems based on two different front-end types: Mel-frequency Cepstral Coefficients (MFCC) and minimum variance distortionless repsonse (MVDR). Confusion network combination (CNC) is applied at the end of each decoding pass.

5.1.4 Spanish Automatic Speech Recognition

We developed a Spanish ASR system that applies the same decoding setup as the described English ASR system. The acoustic models of the four Spanish ASR sub-systems are trained on 140h of Spanish EPPS and Spanish Parliament

	dev06		eval07	
	Spanish	English	Spanish	English
PPL	89	108	89	106
OOV [%]	0.57	1.12	0.83	0.95
WER [%]	8.4	13.9	9.0	12.2

Table 5.1: EPPS English/Spanish ASR system statistics: perplexity (PPL), out-of-vocabulary rate (OOV) and word error rate (WER).

(CORTES) data. Our lowercase pronunciation dictionary has 74.2k . The 4-gram LM is estimated on the Europarl v1[32] Spanish FTEs (25M words), the CORTES texts (44M words) and the EPPS+CORTES (748k words) manual transcriptions. Language model perplexity, out-of-vocabulary rate and case-insensitive WER on our dev06 and eval07 set are listed in Table 5.1.

5.2 English ↔ Spanish Machine Translation

This section describes the English↔Spanish MT system as we developed it for the verbatim translation task of the TC-STAR 2007 evaluation. The author would like to thank all people that contributed to the system build; Jan Niehues for implementing the phrase table smoothing as it was first introduced in [18], Kay Rottmann for providing the part-of-speech based re-ordering scheme [63] and Stephan Vogel for giving valuable advice and insights into statistical phrase-based machine translation.

5.2.1 Word and Phrase Alignment

Phrase-to-phrase translation pairs are extracted by training IBM Model-4 word alignments in both directions, using the GIZA++ toolkit [48], and then extracting phrase pair candidates which are consistent with these alignments, starting from the intersection of both alignments. This is done with the help of phrase model training code provided by University of Edinburgh during the NAACL 2006 Workshop on Statistical Machine Translation [33]. The raw relative fre-

quency estimates found in the phrase translation tables are then smoothed by applying modified Kneser-Ney discounting as explained in [18].

5.2.2 Source-side Word Reordering

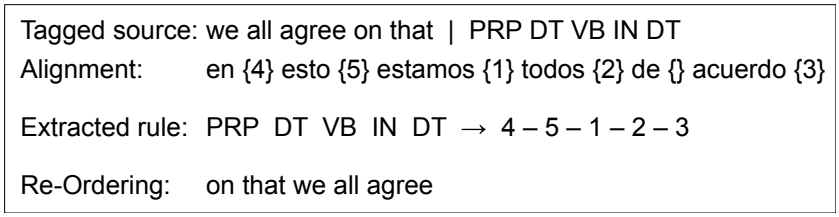

Figure 5.2: Learning part-of-speech reordering rules.

We apply a part-of-speech (POS) based reordering scheme [63] to the source sentences before decoding. For this, we use the GIZA++ alignments and a POS-tagged source side of the training corpus to learn reordering rules that achieve a (locally) monotone alignment. Figure 5.2 shows an example in which a rule is extracted from the POS tags of an English source sentence and its corresponding Spanish GIZA++ alignment. Before translation, we construct lattices for every source sentence. The lattices include the original source sentence along with reorderings that are consistent with the learned rules. All incoming edges of the lattice are annotated with distortion model scores that are dependent on the relative frequency of the learned rules. We refer the reader to [63] for an in depth discussion on how these model scores are computed. Figure 5.3 gives an example of such a lattice. In the subsequent lattice decoding step, we apply either monotone decoding or decoding with a reduced local reordering window, typically of size 2.

5.2.3 Decoder and Minimum Error Rate Training

The ISL beam search decoder [78] combines all the different model scores to find the best translation hypothesis. The presented system applies following models:

- The translation model, i.e. the phrase-to-phrase translations extracted from the bilingual corpus, annotated with four translation model scores. These

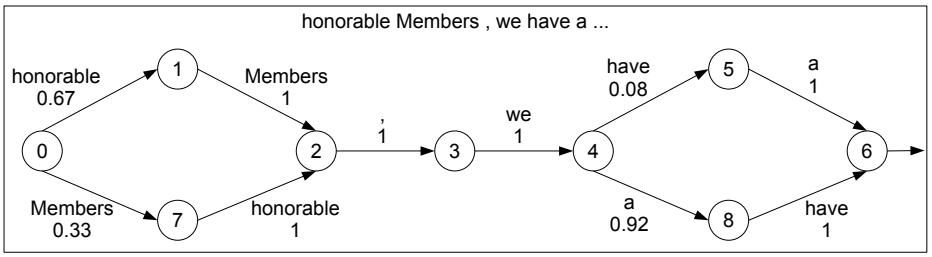

Figure 5.3: Encoding source side reorderings in a lattice structure.

four scores are the smoothed forward and backward phrase translation probabilities and the forward and backward lexical weights.

- A 4-gram LM. The SRI language model toolkit is used to train the LM.
- A 6-gram suffix array LM [84].
- An internal word reordering model. This internal reordering model assigns higher costs to longer distance reordering.
- Simple word and phrase count models. The former is essentially used to compensate for the tendency of the LM to prefer shorter translations, while the latter can give preference to longer phrases, potentially improving fluency.

The decoding process itself is organized in two stages. First, all available word and phrase translations are found and inserted into a so-called translation lattice. Then the best combination of these partial translations is found by doing a best path search through the translation lattice, where we also allow for word reorderings within a predefined local reordering window.

To optimize the system towards a maximal BLEU score, we use minimum error rate (MER) training as described in [46]. For each model weight, MER applies a multi-linear search on the development set n-best list produced by the system. Due to the limited numbers of translations in the n-best list, these new model weights are sub-optimal. To compensate for this, a new full translation is done. The resulting new n-best list is then merged with the old n-best list and

the optimization process is repeated. Typically, the translation quality converges after three iterations.

5.2.4 Training Data Normalization and Statistics

For training, we use the sentence-aligned Europarl corpus v2 [32] combined with the TC-STAR sentence-aligned EPPS parallel text corpus provided by RWTH Aachen [24]. Both corpora are based on EPPS final text editions crawled from the web site of the European Parliament. The resulting parallel text corpus comprises FTEs from April 1996 to May 2006. For a minimal mismatch between training data and source language input to the final MT system, we apply an extensive, automatic pre-processing to the training corpus and the source language input. This pre-processing relies mostly on hand-written rules and includes:

- A data driven true-casing of words. Words at sentence-begin are cased in the same way as they should be cased within a sentence.
- Tokenization of punctuation marks.
- Removal of 'noisy' sections. This includes for example the removal of document references and poorly sentence-aligned sections.
- Expansion of abbreviations and the conversion of numbers and dates to their spoken form.

Further, we remove all bi-lingual training sentences that include more than 80 words on the source or target side. Detailed statistics for our pre-processed training corpus are shown in Table 5.2.

5.2.5 Translation Performance

While our English↔Spanish MT system directly targets the TC-STAR verbatim and spoken language translation task, we also participated with the system in the FTE task (for a more detailed description of the separate tasks refer to Section 4.3). For the FTE task, we include a post-processing step which maps the verbatim-like translation output back to a more FTE-like format. In the

	English	Spanish
sentence pairs	1219415	
unique sent. pairs	1190315	
sentence length	27.3	28.5
words/tokens	33.2 M	34.8 M
proper words	29.8 M	31.4 M
vocabulary	94.2 K	135.6 K

Table 5.2: Training corpus statistics.

following, we present the translation results (NIST BLEU metric, case-sensitive) achieved with our system in the TC-STAR 2007 evaluation. For all three translation tasks, two manual reference translations with proper punctuation are provided. The English and Spanish development and evaluation sets introduced in Section 4.4 are based on the TC-STAR 2007 development and evaluation set for the verbatim task. The references presented in the TC-STAR 2007 development and evaluation set for the ST task are identical to the references of the verbatim task. The official source input to the ST task was generated by combining the ASR system outputs of the individual TC-STAR participating sides using ROVER [17]. The lowercase WER of this source input is 6.9%. The official source input was enriched with punctuation marks that were automatically inserted in a post-processing step after rovering. The translation results of our system (University Karlsruhe, UKA), along with the anonymized results of the best competing systems of other TC-STAR participants, are depicted in Figures 5.4, 5.5 and 5.6.

With the exception for the English→Spanish spoken language translation task, we achieve highly competitive translation results with our MT system. For the SLT task, we decided to rely on the punctuation marks and the sentence segmentation as provided in the official evaluation source input. A post-evaluation error analysis indicated that the performance drop we experienced was a direct result of this decision, as the sentence segmentation and punctuation recovery applied to the English ASR hypotheses of the official evaluation set yielded only

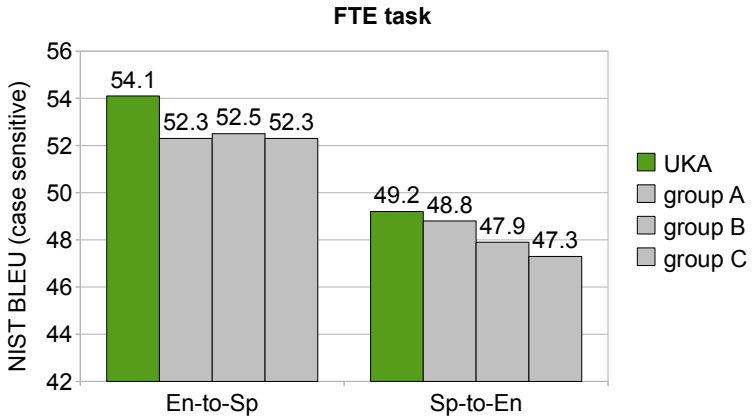

Figure 5.4: Official results for the final text edition (FTE) task of the TC-STAR evaluation 2007.

a very low performance. This result prompted us to develop a more sophisticated sentence segmentation and punctuation recovery approach for spoken language translation, as it is described in the following Chapter 6.

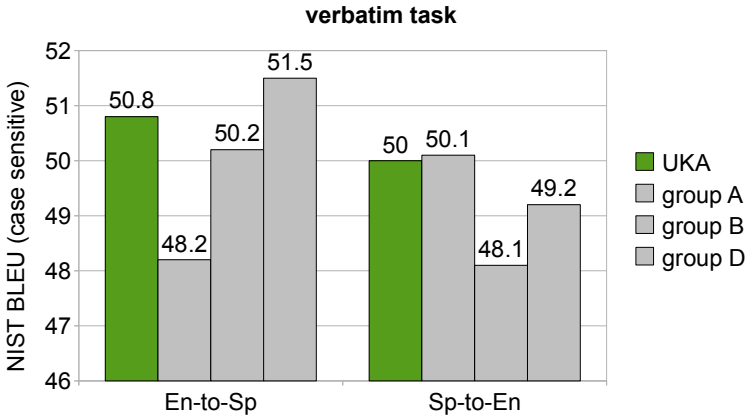

Figure 5.5: Official results for the verbatim task of the TC-STAR evaluation 2007.

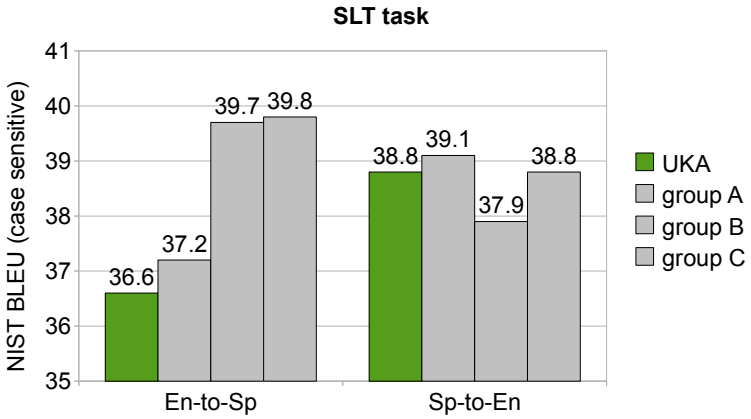

Figure 5.6: Official results for the spoken language translation (SLT) task of the TC-STAR evaluation 2007.

6

Sentence Segmentation and Punctuation Recovery in Speech Translation

Most machine translation training data consists of written text corpora with well-formed sentences and correct punctuation. ASR output, on the other hand, consists of non-sentence like chunks (utterances) of non-punctuated hypotheses. Recognition errors and speech disfluencies like fillers, repetitions and corrections further increase the mismatch between ASR output and MT training data. This severe data mismatch leads to a suboptimal translation performance when applying MT directly to ASR output in the context of spoken language translation. One important step to reduce the given data mismatch is to re-segment ASR hypotheses into sentence-like units before performing translation. Punctuation recovery in ASR output, punctuation removal in MT training data or a mixed approach with a selective punctuation restoration/removal can further reduce the data mismatch. On top of improving translation accuracy, sentence segmentation and punctuation recovery can significantly increase the readability of ST system output. In this chapter, we introduce a combined approach for sentence segmentation and punctuation recovery [52] that uses a decision tree based sentence segmentation system and modified phrase tables. We develop our approach on the basis of our EPPS verbatim MT system, as it was applied to the TC-STAR 2007 English→Spanish spoken language translation task. Further, we successfully

port the developed approach to two additional language pairs: Arabic→English and Chinese→English.

6.1 Related Work

Finding sentence-like units and introducing punctuation in automatic speech recognition output has seen tremendous attention in the past years [12; 25; 28; 39; 69]. An excellent overview on such past work is given by Yang Liu in [39]. Yang Liu summarizes that "previous work has shown that lexical cues are a valuable knowledge source for determining punctuation roles and detecting sentence boundaries, and that prosody provides additional important information for spoken language processing. Useful prosodic features include pause, word lengthening, and pitch patterns. Past experiments also show that detecting sentence boundaries is relatively easier than reliably determining sentence subtypes or sentence-internal breaks (e.g., commas)".

One main motivation for detecting sentence boundaries and introducing punctuation in ASR output is to enhance the readability of the automatic transcriptions. Another main motivation is to "aid downstream language processing tools, which typically expect sentence-like segments" [39]. However, as noted by Rao et al. [26], past work has simply focused on spotting sentence boundaries as defined by humans, mainly ignoring the downstream language processing applied to the ASR output. In this work, we specifically examine sentence segmentation and punctuation recovery in the context of machine translation applied to ASR output of spoken language, similar to work presented in [2; 26]. Both, [2; 26] investigate sentence segmentation to improve spoken language translation accuracy. Rao et al. [26] report improvements in translation accuracy by introducing non-punctuated intra-sentence segments before translation. Al-Onaizan and Mangu [2] investigate the recovery of punctuation during translation, by applying translation phrase tables that do not include punctuation on the source side, but on the target side.

6.2 Experimental Setup

6.2.1 Data & Scoring

Our experiments on English→Spanish ST use the post-processed, ROVERed ASR hypotheses files, as they were provided during the TC-STAR 2006 and 2007 evaluations for the spoken language translation (SLT) task. Both evaluation sets exhibit a case-insensitive WER of 6.9%. For our experiments on Arabic→English and Chinese→English ST, we extract two data sets per language pair from the shadow data included in the ROSETTA team ASR output of the GALE [21] 2007 evaluation. Table 6.1 list the data statistics of the used development and evaluation sets.

	English		**Arabic**		**Chinese**	
	dev (eval06)	eval (eval07)	dev	eval	dev	eval
words/chars [k]	30	26	8	9	23	8
WER/CER [%]	6.9	6.9	12.1	21.7	10.5	17.1

Table 6.1: Sentence segmentation and punctation recovery: data statistics, including word error rate (WER) for Arabic and English and character error rate (CER) for Chinese.

We measure the success of our approaches for sentence segmentation and punctuation recovery in terms of an end-to-end translation performance using BLEU metric on human translation references that include proper punctuation. For English→Spanish, two case-sensitive translation references are used and we report case-sensitive BLEU scores. The translation references for Arabic→English and Chinese→English are lowercase and comprise only one reference per source sentence.

6.2.2 MT Systems

The En→Sp spoken language translation experiments reported in this chapter are based on the MT system described in Section 5.2. For Ar→En and Ch→En

spoken language translation, we apply statistical phrase-based MT systems as they were trained at the Interactive Systems Labs for the GALE 2007 evaluation. The training setup of these systems is similar to the training setup described in Section 5.2. In contrast to the En→Sp system, no reordering based on part-of-speech tags is applied to the source sentences prior to translation. Instead, a simple word reordering which assigns higher costs to longer distance reorderings is used. The Ar→En system uses a reordering window of four words and Ch→En system uses a reordering window of two words.

6.2.3 Baseline Segmentation

The baseline segmentation of the Arabic and Chinese ASR hypotheses is identical to the automatic speech utterance segmentation that was inferred prior to ASR via voice-activity detection. The baseline segmentation of the English ASR hypotheses found in eval06 and eval07 is based on the periods that are included in these sets. Punctuation marks (period, comma) were inserted in these sets before distributing them to the TC-STAR participants. The insertion of punctuation marks relied on University Karlsruhe's punctuation recovery system, as we originally developed it for the TC-STAR 2006 evaluation. This simple punctuation recovery system is based on local language model context and pause duration between words. A period is explicitly inserted if the pause duration p at a given word boundary B_i is bigger than 0.7 seconds. For $0.03s < p \leq 0.7s$, the insertion of a period or a comma is determined by the local LM context $w_{i-2}B_{i-1}w_iB_iw_{i+1}B_{i+1}w_{i+2}$, with B_{i-1} being the boundary / punctuation mark type estimated in the previous step $i-1$. For B_{i+1}, all possible punctuation marks are considered.

6.3 Comma Recovery via Modified Phrase Tables

Punctuation recovery for SLT can either be performed before, after or implicitly during translation. Punctuation recovery before translation can rely on acoustic features (prosody, pause duration, etc.) that are extracted from the speech

signal. Our experiments indicate that such features are especially of value for the recovery of periods. However, introducing punctuation in the source may have the disadvantage to degrade translation performance due to punctuation errors. A faulty punctuation may split up long source phrase matches into two or more short phase matches. Punctuation recovery during or after translation does not suffer from this problem, as non-punctuated source phrases guarantee a maximal match. Further, these approaches are more strongly influenced by the target language model, which possibly results in a more 'natural' punctuation of the translation output. Our experiments indicate that these considerations are especially important in the context of comma recovery. Punctuation recovery during translation can be realized by removing punctuation from the source side of the parallel MT training data while retaining punctuation on the target side. However, Al-Onaizan and Mangu [2] point out that this will likely degrade word alignment accuracy as it may not be possible to correctly align target punctuation. Instead, they propose to remove punctuation from the source phrases in the phrase table of a phrase-based statistical MT system, while retaining punctuation in the target phrases.

We explicitly distinguish between two separate tasks; sentence segmentation and punctuation recovery. A sentence segment constitutes the unit presented to MT; MT processes each unit independently, one after another. A sentence segment may include one or multiple punctuation marks (period, comma)—or none at all. By explicitly separating sentence segmentation from punctuation restoration, it is possible to fully explore punctuation recovery before, after or during translation.

We examine the effect of performing punctuation recovery before and/or during translation by comparing English→Spanish SLT performance when (1) retaining periods and commas as presented in eval06 and eval07, (2) removing all punctuation marks and applying implicit recovery of periods and commas during translation, and (3) retaining periods in the ASR output and recovering commas during translation. For (2) and (3) we initially train phrase-tables with punctuation and then remove punctuation from the source side as described in [2]. Table

6.2 shows the achieved SLT performance in BLEU. Inserting full stops prior to MT, on the *source* side at each segment end and recovering commas implicitly during translation using a modified *phrase table* obtains the best translation performance. For Arabic and Chinese we observe that inserting periods at each segment end leads to an improved translation performance, even for the baseline speech/non-speech audio segmentation. This concurs with results presented in [2]. Table 6.3 shows the influence of comma recovery via modified phrase tables on Ar→En and Ch→En SLT performance. For all subsequent experiments reported in this chapter we include a period at the end of each source segment and we apply modified phrase tables for comma recovery.

period	**comma**	**dev**	**eval**
source	*source*	38.7	36.6
phrase table	*phrase table*	39.2	38.3
source	*phrase table*	40.2	39.0

Table 6.2: En→Sp BLEU scores for different punctuation recovery schemes: source side vs. modified tables.

	period	**comma**	**dev**	**eval**
Ar→En	*source*	-	19.5	13.5
	source	*phrase table*	21.3	15.3
Ch→En	*source*	-	8.3	8.7
	source	*phrase table*	8.8	10.1

Table 6.3: Ar→En and Ch→En BLEU scores without comma recovery and with comma recovery using modified phrase tables.

6.4 Decision Tree Based Sentence Segmentation

To improve sentence segmentation compared to the baseline segmentation described in Section 6.2.3, we develop a decision tree based sentence segmentation architecture that uses multiple word boundary features. We use J.R. Qinlans

C4.5 induction system [60] for decision tree training and rule extraction. We create the necessary training examples by automatically aligning ASR hypotheses to their reference transcriptions, using RWTH Aachen's multiple word error rate (mWER) segmentation tool [42]. Such mWER segmented ASR hypotheses observe the segmentation of the manually created reference transcriptions, while still including typical ASR transcription errors. This guarantees a minimal mismatch between training and evaluation data. Using different word boundary feature combinations, we select the decision tree and feature set combination that yields the highest F-measure in regards to human segmentation on a development set.

For English, we train the decision tree on the English dev (eval06) ASR hypotheses. The final feature set combination consists of word duration of the word preceding the current boundary, pause duration and LM probabilities for comma and full stop insertion (with the same local LM context as described in Section 6.2.3). For Chinese, we train the decision tree on shadow data included in the ROSETTA team 2007 ASR dry-run. This data consists of 6 shows from the GALE 2006 development set and of the second half of the GALE 2007 development set. For Arabic, we use 4 shows from the BNAD05 data set. For both languages, the final feature set combination consists of pause and word duration as well as LM probabilities for full stop insertion. For Arabic, we also include prosody based features. Specifically, we encode pitch information by combining pitch and delta pitch values in the vicinity of 700 milliseconds of the candidate boundary. We also included the signal power values in the same region as well as total signal power on either side of the boundary. As high dimensional features cause data sparsity problems and result in over-fitting of the decision tree, we reduce the dimensionality by training a support vector machine based classifier on these features. We then use the scores of the support vector machine classifier as features within the decision tree. We considered the same prosody based features for English and Chinese sentence segmentation. However, for these languages we did not observe any improvements in terms of F-Measure by adding the prosody based features to our standard feature set. Table 6.4 compares the

F-Measures of the baseline segmentation and the decision tree based segmentation. Table 6.5 lists the BLEU scores of the end-to-end system, depending on the used sentence segmentation. For all three language pairs, the decision tree based sentence segmentation achieves consistently higher BLEU scores than the baseline segmentation. In addition to the results for the baseline segmentation and the decision tree based segmentation, we also list the BLEU scores achieved when using mWER sentence segmentation (the sentence segmentation of the transcription references). The results show that human style sentence segmentation results in improved automatic translation performance.

	segmentation	**dev**	**eval**
English	baseline	54.79	52.48
	decision tree	65.97	62.14
Arabic	baseline	33.89	37.50
	decision tree	40.97	43.41
Chinese	baseline	30.75	31.59
	decision tree	59.16	53.38

Table 6.4: F-Measures; baseline segmentation vs. decision tree based segmentation.

6.5 Phrasal and Target LM Context for Source Side Sentence Segmentation

Different source side sentence segmentations lead to different source phrase matches and different target side language model histories during translation. Possible word and phrase re-orderings during translation are also affected. For a better integration of source sentence segmentation and phrase based MT, we experiment with features to incorporate phrasal and target language model context during source sentence segmentation. To infer such features, we apply knowledge derived from the translation beam-search lattice as it is constructed during decoding (we

	segmentation	dev	eval
En→Sp	baseline	40.2	39.0
	decision tree	40.3	39.5
	mWER	41.4	41.3
Ar→En	baseline	21.3	15.3
	decision tree	21.4	15.5
	mWER	21.8	16.0
Ch→En	baseline	8.8	10.1
	decision tree	8.9	10.7
	mWER	9.4	11.0

Table 6.5: BLEU scores for different segmentations: baseline segmentation, decision tree based segmentation and multiple word error rate segmentation.

refer the reader to Section 3.3.1 for a more detailed description of the translation lattice). The motivation is not to break up source phrases that are valuable for MT and also to pay attention to the target LM context during sentence segmentation.

To compute a score indicating if phrasal context or target LM context is jeopardized when segmenting at a given word boundary, we apply a sliding window of 24 words with a step size of 6 words on the ASR output. For each step, we translate the 24 word sentence and compute two probabilities, $phrSP$ and $tbiSP$, for each of the 11 word boundaries between the innermost 12 words. These two probabilities are computed from the translation lattice used by our beam-search decoder. The edges in this lattice correspond to source words and phrases (together with their translation) and the nodes to the boundaries between these words and phrases. The phrasal split-point probability $phrSP$ for a given word boundary is computed as the number of paths going over its corresponding node divided by the number of paths visiting its node. We consider only the n-best pathes, i.e. the n-best translations. A phrasal split-point probability of one indicates that the word boundary is always seen between two source phrases in the n-best translations. Introducing a segment boundary at such a point should

therefore not negatively affect possible phrase matches during translation. The target LM split-point probability $tbiSP$ is computed only for word boundaries with $phrSP > 0$ and is based on bi-gram probabilities. For all m word boundaries that are found to lie between two phrases, the target LM probability tbi of the bi-gram formed by the last word of the left source phrase and the first word of the right source phrase is computed. If the target LM does not include an according bi-gram, a bi-gram probability of 0 is assumed. $tbiSP$ is defined as: $tbiSP = 1 - (\sum^{m} tbi)/m$.

We analyze the correlation of the phrasal split-point probability $phrSP$ with actual sentence boundaries. We compute $phrSP$ for all word boundaries found in the human transcriptions of the English dev set using the 100-best translations. We then select six split-point probability ranges. For each range, we compute the percentage of sentence boundaries compared to the absolute number of boundaries within the range. Figure 6.1 shows the result. While a high phrasal split-point probability does not necessarily predict a sentence boundary, a low phrasal split-point probability seems to be a strong indicator of a non-sentence boundary. However, augmenting our decision tree based sentence segmentation with $phrSP$ as an additional feature did not lead to any significant improvements.

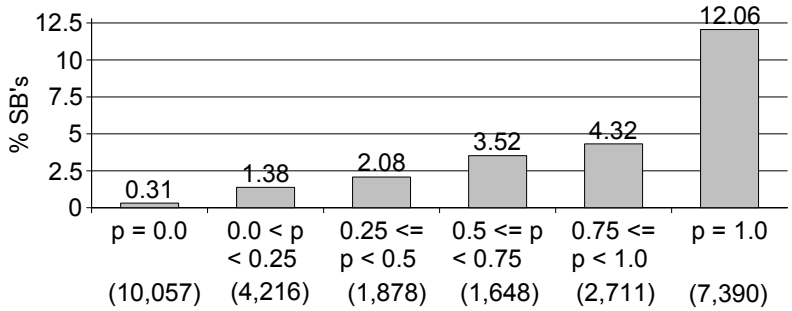

Figure 6.1: Percentage of sentence boundaries compared to the absolute number of boundaries (words) within different phrasal split-point probability ranges p.

We repeat a similar experiment for different target LM split-point probability $tbiSP$ ranges r with $0.0 < r <= 1$. The ranges are selected in a way that

each range contains approximately 1600 boundaries (words). The continuous line in Figure 6.2 shows the percentage of sentence boundaries included in the different ranges. While there seems to be no clear correlation between target LM split-point probability and human sentence boundaries, an increase of included sentence boundaries around range $r4$ can be observed. When plotting the percentage of bi-grams that include a comma in the same graph (dotted line), we see that the same region has a high percentage of bi-grams including commas. Phrase boundaries that are connected with a comma therefore seem to correlate stronger with human sentence boundaries. This coincides with the intuition that distinguishing between a comma boundary and a full stop boundary when transcribing spoken language is often times up to interpretation and dependent on the style of the individual human transcriber.

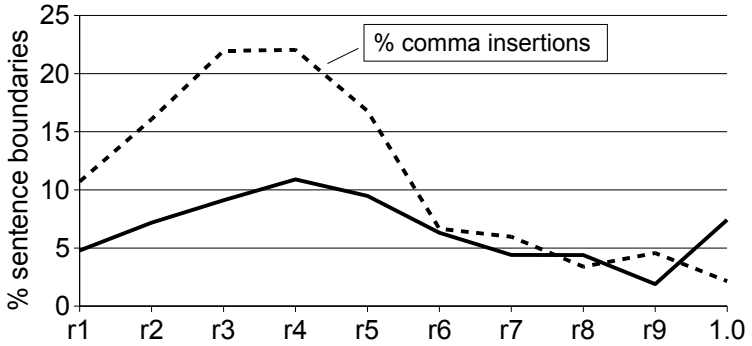

Figure 6.2: Percentage of included sentence boundaries (continuous line) for different target language model split-point probability ranges r.

6.6 Chapter Summary & Discussion

We described our sentence segmentation and punctuation recovery scheme for spoken language translation. By applying modified phrase tables for implicit target side comma recovery during translation and by introducing a decision tree based sentence segmentation for insertion of full stops on the source side,

we significantly improved translation performance on three language pairs. Results in BLEU are summarized in Table 6.6. Furthermore, we investigated two novel features indicating if phrasal context and target language model context is jeopardized when segmenting at a given source word boundary. However, no additional gains in end-to-end translation performance could be observed with these features.

	En→Sp	**En→Sp**	**En→Sp**
baseline	36.6	8.7	13.5
combined approach	39.5	10.7	15.5

Table 6.6: Improved spoken language translation performance, measured in BLEU, by applying our combined sentence segmentation and punctuation recovery scheme.

PART II

Audio Recordings of Human Interpretation as a Novel Resource for Speech Translation System Development

7

Interpretation: A Data Resource for Speech Translation?

7.1 Terminology

Translation refers to the transfer of meaning from source language text to target language text, with time and access to resources as dictionaries, phrase books, et cetera. *Interpretation* (of speech) refers to the transfer of meaning from source language speech to target language speech, either simultaneously, while the source language speaker continuously speaks, or consecutively, with source language speaker and interpreter taking turns. We define the term *parallel speech (pSp)* as speech of a source language speaker together with the target language speech of an interpreter. Parallel speech therefore always refers to either simultaneous interpretation (SI) or consecutive interpretation (CI). It specifically excludes speech of translators as it for example occurs in the context of automatic dictation systems for translators.

7.2 The Nature of Interpretation

Figure 7.1 gives an example for (manually transcribed) parallel speech as it occurs in SI and CI. The figure also provides a manual translation of the non-English

parallel speech. Comparing the provided interpretation and translation, significant differences become immediately apparent. To understand why and how interpretation differs from translation, it is necessary to take a closer look at the strategies applied by interpreters.

Simultaneous Interpretation	Consecutive Interpretation
SPANISH UTTERANCE: “trataremos de que todo el personal tenga” *TRANSLATION:* *“we shall try that all the staff will get”* PARALLEL SPEECH: “... in addition to that we are going to try to make sure that members of staff from different members states of the european union will be granted an equal status ...”	ENGLISH UTTERANCE: “okay and what is the importance of this gas station” PARALLEL SPEECH: "بېخي صحىح ده د دې په هکله تا څه غوښتل چې زما سره وواېي" *TRANSLATION:* *“it is okay - what do you want to tell me about this”*

Figure 7.1: Interpretation (parallel speech) versus translation.

The strategy of ‘dropping form’ is one of the main reasons why interpretation and translation differ strongly, even if the interpreter conveys all elements of meaning. Dropping form refers to the fact that interpreters immediately and deliberately discard the wording and retention of the mental representation of the message [68]. Only by discarding the words, sentence structure, etc., interpreters—in SI as well as in CI—are able to concentrate on the meaning of the message and its reformulation in the target language [64]. The reason for this lies within the limitations of the human short-term memory. Only up to six or seven items can be retained in short-term memory, and only if we give all of our attention to them [70].

In the case of SI, the difference to translation is also strongly influenced by special strategies interpreters have to apply to keep pace with the source language speaker. These strategies include anticipation-strategies [7] and compen-

satory strategies [1]. For example, interpreters anticipate a final verb or syntactic construction before the source language speaker has uttered the corresponding constituent. The interpreter confirms this anticipation or corrects it when he receives the missing information. The use of open-ended sentences that enable the interpreter to postpone the moment when the verb must be produced is another anticipation-strategy. Compensatory strategies include skipping, approximation, filtering, comprehension omission and substitution. Corrections of previous interpretation errors as well as fatigue and stress also negatively affect SI quality. It is important to note that SI can result in a significant loss of information. Experiments reported during the course of the TC-STAR project [43] suggest that the information loss for English-to-Spanish SI as provided during EPPS amounts to approximately 9%. This number is based on a test set of comprehension questions created from the English speech and the respective amount of answers that cannot be found in the Spanish SI. Further, it is reported in [43] that the effective information loss is with 29% significantly higher. This effective information loss results from the combination of missing information and the difficulty of human evaluators to follow the flow of—often syntactically misformed—interpreter speech. In the reported evaluation scenario, the human evaluators are allowed to listen to the recorded interpreter speech twice and they could interrupt the playback to write down their answers.

In CI, interpreters face less severe time constraints, resulting in more accurate, equivalent, and complete interpretations [64]. However, the less severe time constraints in CI can also contribute to the differences between interpretation and translation, as "interpreters also elaborate and change information and they do not only convey all elements of meaning, but also the intentions and feelings of the source speaker" [36]. We speculate that these effects are more prevalent in CI than in SI, as CI scenarios tend to be more personal and the interpreter has more time to elaborate.

7.3 Interpretation and Automatic (Speech) Translation

To measure the performance of automatic translation of speech or text we apply the widely adopted BLEU metric, which, on a scale from 0–100, compares MT output to one or more human reference translations based on n-gram comparison. Like all automatic evaluations, BLEU is not able to determine if a given MT output correctly translated all meaning, but can only determine how 'similar' the output is to given reference translations. As described in detail in Section 7.2, we cannot expect interpretation to be 'similar' to translation, even if an interpretation captures all meaning.

To underline this point, we compute the BLEU score of the manual transcription of Spanish-to-English (Sp→En) and English-to-Spanish (En→Sp) SI speech, as present in our dev05 set (for a description of dev05, refer to Section 4.4). Given two reference translations, only 14.2 BLEU for Sp→En and 18.2 BLEU En→Sp are achieved. This compares to scores of above 40 BLEU points, achieved by state-of-the art spoken language translation systems (trained on approximately 100h of transcribed speech and 30+ million translated words), as developed within the European project TC-STAR.

7.3.1 A Hypothesis

Despite this low 'translation' performance of interpreters (measured in BLEU) we argue that interpretation has the potential to be a valuable resource for automatic translation of text and speech—even if we only measure its value in terms of machine evaluation metrics, like BLEU score. Apart from the fact that an interpretation should, ideally, represent the same elements of meaning as a respective translation, its potential value is already indicated by the fact that the above computed BLEU scores for SI are not zero. In other words, we believe that the existing n-gram matches between interpretation and translation can already be of value to improve automatic translation performance in the context of MT

and ST. Figure 7.2 gives an example for the n-gram matches that can be found between interpretation and translation.

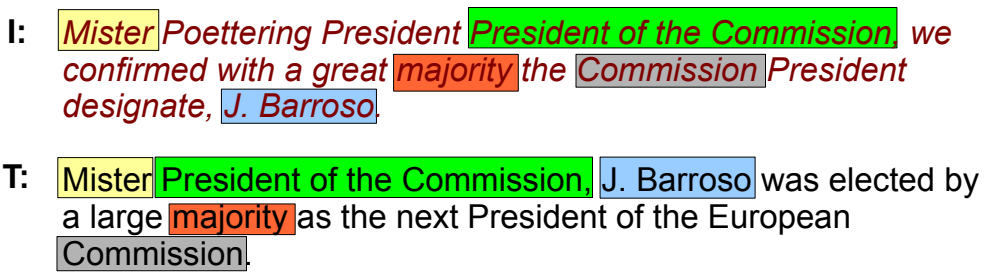

Figure 7.2: N-gram matches between interpretation (I) and translation (T).

By completely ignoring the aspect of meaning and relying on the same core ideas of statistical phrase-based MT, we can simply look at interpretation as some form of 'noisy' translation, that still includes valuable word-to-word or phrase-to-phrase translations. Having this concept in mind, we can simply state that the amount of 'noise' depends on how far the interpreter deviated in his interpretation from the original wording of the source speech. Given the fact that (a) an interpretation should ideally convey the same elements of meaning as a translation does, and (b) there are only 'so many ways' one can express the same concept in a given language, we could go even further and speculate that to cover the same phrase-to-phrase translations found in a specific bi-lingual translation corpus, it is simply necessary to acquire a larger interpretation corpus (on the same topic). A statistical phrase-based MT system, based solely on these matching phrase-to-phrase translations, would necessarily yield a very similar translation performance as a system trained on the smaller bi-lingual translation corpus.

7.3.2 Prospective Use of Interpretation Data

Speech translation combines two technologies, ASR and MT. Accordingly, we can identify two separate goals when trying to exploit interpretation as an additional data resource for ST. From an ASR point of view, we want to improve transcription performance while from a MT point of view, we desire to improve

translation performance. Both goals contribute to an improved end-to-end system performance. Considering these two goals, we can identify different prospective use cases for interpretation data.

In a scenario where we seek to automatically transcribe the speech of a source language speaker (in language A) and the parallel speech of an interpreter (in language B), we should be able to improve the recognition performance for language B, by biasing ASR_B with knowledge derived from the source speech. After all, the interpreter should deliver the same elements of meaning expressed by the source speaker. The speech in language A may therefore serve as a predictor of what the interpreter is going to say. To accomplish such a biasing of ASR_B, we can for example apply A→B speech translation and use the resulting translation to adapt the language model of ASR_B. In the same manner, we can bias ASR_A.

The described approach is based on ideas first presented by Brown et al. [9] and Dymetman et al. [14] for improving the performance of dictation systems for human translators in the context of machine aided translation. While pervious works only considered biasing target language dictation systems with knowledge extracted from source language documents, we applied the described approach in [53; 54] for the first time to extract knowledge from source language speech. However, our experiments presented in [53; 54] only considered read-speech, with source and target language speaker reading from a travel-domain parallel text corpus of sentence-aligned translations. In contrast to this rather artificial task, we consider in this thesis parallel speech audio, as it occurs in the form of simultaneous interpretation within the European Parliament. In [53; 54], we also exploited the ASR_A (ASR_B) transcriptions to bias the B→A (A→B) machine translation system, leading to an improved automatic translation performance. This improved translations were then used in an iterative manner to further improve the transcription performance of ASR_A and ASR_B.

The technique of biasing ASR (MT) with parallel speech promises an improved transcription (translation) performance on parallel speech only. However, an automatically transcribed parallel speech corpus may be valuable as additional

training data that helps to improve the **general** performance of the statistical models involved in speech translation. In other words, it may be possible to train automatic speech translation from parallel speech and apply such a system in situations where no interpreter is available. The automatic transcriptions, together with the original audio, may serve as additional acoustic model training data. Further, the automatic transcriptions may be used as additional language model training data. Finally, following the argumentation given in Section 7.3.1, the automatically transcribed parallel speech may be tied together in a parallel text corpus suitable for translation model training.

7.3.3 Automatic Interpretation

So far, we only discussed the prospective value of parallel speech (interpretation) as a data resource for automatic translation of speech or text. In fact, even for the prospective application case of easing the simultaneous interpretation effort of EPPS with automatic methods (Figure 4.1), we speak of automatic real-time speech-to-speech (S2S) *translation*. The question arises if it is not desirable to achieve automatic interpretation or, in other words, to simulate the techniques and strategies applied by human interpreters.

In Section 7.2, we listed results regarding human simultaneous interpretation performance, as reported in [43]. These results indicate that simultaneous interpretation, as presented during EPPS, can result in a significant loss of information. Only 74% of content questions, created from an English politician speech, could be answered by human judges after the judges were allowed to listen twice to a Spanish interpretation of the English speech. This significant information loss is a direct result of the strategies applied by human interpreters. Therefore, we argue that it is not desirable to simulate the techniques/strategies applied by human interpreters in the form of automatic interpretation systems.

As described in Section 7.2, Mostefa et al. [43] further report that the actual missing information in the Spanish interpreter speech amounts only to approximately 9% and that the total amount of information loss results from an inability

of human judges to extract all information from interpreter speech. Mostefa et al. [43] repeated the same experiment, replacing the Spanish interpreter speech with the system output of the TC-STAR 2007 end-to-end system. Figure 7.3 depicts the overall results. Only 64% of the content questions could be answered, with the amount of total information not included in the system output being approximately 11%. The performance of the automatic speech-to-speech translation system is surprisingly close to the performance of the human interpreter. Due to human cognitive limitations, simultaneous interpretation will always suffer from a significant information loss. However, automatic S2S translation continues to be subject to tremendous research activity and has access to ever faster processors and larger amounts of memory. Therefore, the question arises, if real-time S2S translation—combined with methods of delivering the automatic translation in an effective way, e.g. by applying summarization techniques—may one day outperform human simultaneous interpretation performance. The very first open-domain, real-time S2S translation system for lectures and speeches was presented in late 2005 at the Interactive Systems Laboratories (interACT). First steps towards the development of this system were already taken as early as 1998 by developing automatic transcription (and browsing) systems for meetings and lectures [62; 79].

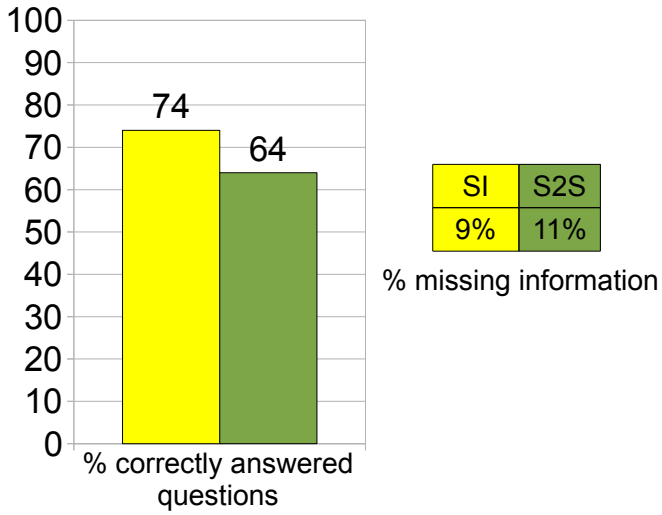

Figure 7.3: TC-STAR comprehension evaluation; simultaneous interpretation (SI) vs. speech-to-speech (S2S) translation.

7. INTERPRETATION: A DATA RESOURCE FOR SPEECH TRANSLATION?

8

Interpretation as an Auxiliary Information Source

In Section 7.3.2 we discussed several prospective benefits of interpretation data for spoken language translation. For example, we argued that approaches previously applied in dictation systems for human translators should be transferable to the simultaneous interpretations provided during EPPS and should therefore be beneficial to the EPPS verbatim transcription and translation task. In this chapter, we explore such approaches. Specifically, we seek to improve automatic transcription and automatic translation applied—in an offline fashion—to the parallel speech of politician and simultaneous interpreter, by exploiting the parallel information given in the respective other language stream. In Section 7.3.2 we argued that a (well) transcribed parallel speech corpus may be of value for training ST models. The approaches introduced in this chapter for improving ASR of parallel speech will be applied to ST model training in Chapter 11.

8.1 Experimental Setup

8.1.1 Data and Scoring

The experiments presented in this chapter are based on the EPPS dev05 and dtest05 sets, as described in detail in Section 4.4. Each of these sets represents an English/Spanish parallel speech corpus. Further, for each of these sets, the

speech in the parallel English and Spanish audio tracks switches between politician speech and interpreter speech. We do not explicitly distinguish between politician or interpreter speech in the following. It is possible that some time segments of the parallel English and Spanish audio contain only interpreter speech on both channels, and no politician speech. In such a situation, the politician that took the floor in the European Parliament spoke in a language different from English or Spanish.

To evaluate ASR performance, we compute case-sensitive WER for Spanish ASR and case-insensitive WER for English ASR. To evaluate automatic translation performance, we rely on two non-punctuated, case-sensitive translation references for NIST BLEU score computation. In addition to the BLEU scores based on these two reference translations, we also provide BLEU scores based on the reference transcription of respective parallel speech. To distinguish between these two BLEU scores, we provide the latter always in square brackets. At the very end of this chapter, we will also use BLEU metric to express the 'translation' performance of parallel speech. This is accomplished by computing the BLEU score of the pSp reference transcription in respect to the available two reference translations. Figure 8.1 gives a schematic overview on the data input involved when computing these three different BLEU scores and lists the according baseline scores for the translation direction En→Sp on dev05.

8.1.2 Baseline Systems

All systems used in this chapter are based on the EPPS systems described in Chapter 5. The English ASR is identical to the previous described English ASR, while the Spanish ASR differs from the previously described Spanish ASR in its pronunciation dictionary and language model—here, we use case-sensitive versions. We do not use the sentence segmentation and punctuation recovery scheme described in Chapter 6, but we apply MT directly on ASR output. To tailor our MT system to the task of translating non-punctuated ASR hypotheses, we remove source and target side punctuation from the phrase table entries. Since the English ASR produces only lowercased hypotheses, we also lowercase the English

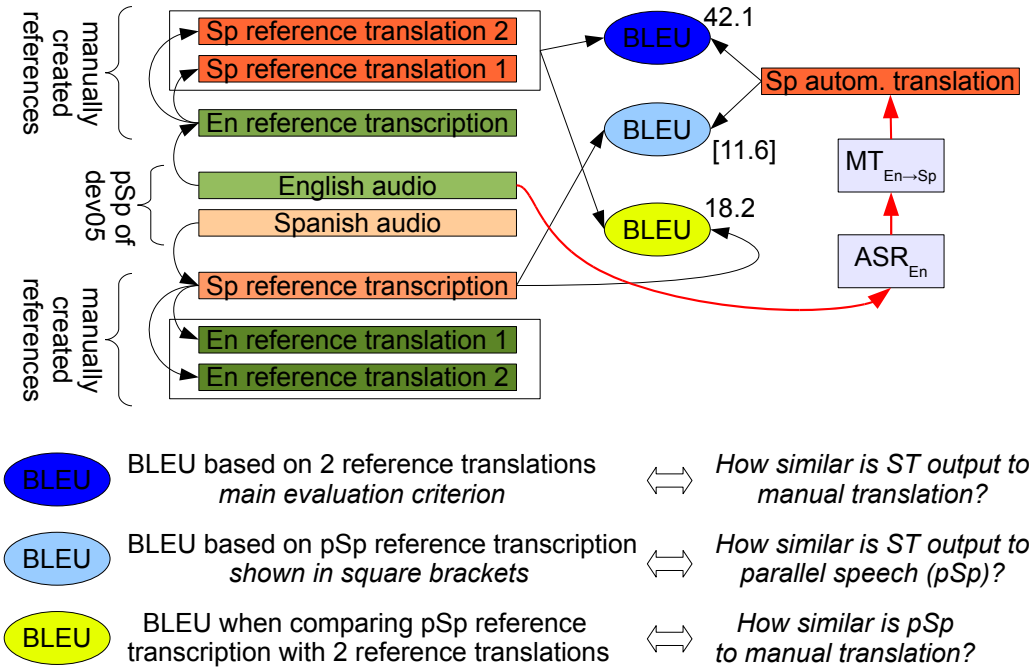

Figure 8.1: Interpreting the provided BLEU scores correctly: data input involved in score computation and example scores for En→Sp on dev05.

phrase table entries for the English→Spanish translation direction. Further, we do not apply the POS-based reordering scheme described in Section 5.2.2, but only rely on the internal word reordering model of the ISL beam search decoder, using a reordering window of 2.

8.1.3 Confusion Network Translation

As discussed in Section 3.4, it is possible to improve speech translation performance by considering multiple ASR hypotheses for MT, instead of applying MT only on the first-best ASR hypotheses. Bertoldi et al. [5] have shown that the translation of ASR confusion networks (CNs) is an efficient integration strategy that improves translation performance. The ISL beam-search decoder does not natively support confusion network translation. In order to process ASR confusion networks, it is necessary to transform the CNs into the native lattice input format of the decoder. Due to decoder constraints, the internal word reorder-

ing model works only on input lattices where all source sentences found within the lattice are of equal length. Confusion networks do show this property, since they assure source-sentences of equal length by introducing special empty words ϵ. To handle ϵ correctly during translation, we further modify the phrase tables of our MT system. From the 10k best paths of each CN, we extract all n-grams that include ϵ. We then extend the phrase tables[1] by duplicating all entries where the source phrase would match an extracted n-gram if the ϵ would not be there. Finally, we replace the original source phrases of these duplicates with the n-gram containing ϵ. Table 8.1 compares the translation performance when translating the reference transcriptions, ASR first-best hypotheses and the confusion networks. Confusion network translation outperforms ASR first-best translation by 0.7 to 0.8 BLEU points on dtest05 in both translation directions. We use confusion network translation for all remaining spoken language translation experiments in this chapter.

	dev05			**dtest05**		
	ref.	1-best	CN	ref.	1-best	CN
En→Sp	45.1	40.3	42.1	44.2	39.8	40.6
Sp→En	54.7	48.6	50.1	52.2	44.8	45.5

Table 8.1: BLEU score for translating reference transcriptions, ASR 1-best hypotheses and ASR confusion networks (CN).

8.2 Biasing Machine Translation

8.2.1 MT Language Model Adaptation

For biasing the machine translation LM with knowledge extracted from parallel target language speech, we mostly rely on target language n-grams extracted from an ASR first-best hypothesis of the target language speech. Specifically, we extract all n-grams with $n \in \{1, 2, 3\}$ from the ASR first-best hypothesis of the

[1]We use individual phrase tables, one per source sentence. These phrase tables are loaded dynamically during decoding.

target language audio snippet that begins 6 seconds before and ends 6 seconds after a source language utterance. We chose this 6 seconds padding of the target speech as we observed that, if the information contained in the source utterance is at all present in the parallel speech, such a padding almost always guarantees that the respective information is present in the parallel speech snippet. In the following, we refer to the extracted n-grams as ASR n-gram hints. Prior to extracting ASR n-gram hints from the target language ASR first-best hypothesis, we remove all words from the hypothesis that have a word confidence of $c < 0.9$. For each source utterance, we dynamically load its respective ASR n-gram hints during MT decoding. Whenever an ASR n-gram hint is observed during decoding, a discount is applied to the LM score (cost) of the current translation hypothesis. In this way, we favor translations that contain ASR n-gram hints. The discount value of each hint is computed as w_n times its logarithmic LM probability. The LM probability of each ASR n-gram is determined by the unadapted LM. The value for w_n is estimated via MER optimization. In addition to this scheme, we penalize all words that are not part of the ASR first-best vocabulary. The factor by which we penalize non-ASR vocabulary words is again estimated via MER optimization. The ASR first-best vocabulary is determined on a 'per session' basis[1], rather than on a per utterance basis. Tables 8.2 and 8.3 list the achieved BLEU scores for the two translation directions. In addition to the BLEU scores computed with two reference translations, we list in brackets the BLEU scores computed with the ASR reference transcription of the respective parallel speech in the target language. For both translation directions we observe consistent gains in translation ['interpretation'] performance, measured in BLEU. In other words, by biasing the MT language model towards parallel speech n-grams, we achieve automatic translations that do not just contain more parallel speech n-gram matches, but also more translation reference n-gram matches. Another observation is that the BLEU scores computed with the parallel speech reference transcription shows a much stronger deviation between dev05 and dtest05 than the BLEU scores computed with the two translation references. This possibly indicates that parallel speech (interpretation) is less consistent than translation.

[1] Dev05 and dtest05 both contain only one parliamentary session each.

	dev05		dtest05	
	ref.	CN	ref.	CN
Baseline	45.1 [11.4]	42.1 [11.6]	44.2 [15.9]	40.6 [14.4]
LM bias	46.2 [13.8]	43.3 [14.4]	44.7 [18.4]	43.0 [18.2]

Table 8.2: En→Sp BLEU scores when biasing the Spanish MT language model with Spanish parallel speech. Results are listed for reference transcriptions (ref.) as input to the MT system and ASR confusion networks (CN) as input to the MT system.

	dev05		dtest05	
	ref.	CN	ref.	CN
Baseline	54.7 [8.9]	50.1 [8.2]	52.2 [19.9]	45.5 [17.0]
LM bias	55.6 [10.4]	51.6 [9.9]	52.9 [24.2]	46.4 [20.9]

Table 8.3: Sp→En BLEU scores when biasing the English MT language model with English parallel speech. Results are listed for reference transcriptions (ref.) as input to the MT system and ASR confusion networks (CN) as input to the MT system.

8.2.2 Translation Model Adaptation

In addition to biasing the MT language model, we also conduct experiments to bias the source sentence specific phrase tables. For this, we extract 'ASR translation phrases' by computing the alignment matrix between the ASR first-best hypotheses of a source utterance and its respective ±6 seconds padded target language audio snippet. In a first iteration, this alignment matrix consists only of word-to-word translation probabilities extracted from the forward and backward IBM4 lexicons that were computed during MT phrase table training. We then estimate for each source word a discrete probability distribution for source-to-target word delays d, with $d \in \{-6, .., 0, .. + 6\}$ seconds. The source-to-target word delay is defined as the distance in seconds between the start time of the

source words and their respective target language translation in the parallel audio. For estimating the discrete probability distribution, we consider only words that are aligned with a high lexical translation probability and that are found within a 60 second window around the current source word. The alignment matrix is then re-estimated, using an interpolation of lexical translation probabilities and the estimated delay alignment probability. In a next step, we introduce binary alignment links. These binary alignment links are computed with the help of a simple algorithm described in [75]. This algorithm allows limited alignment link overlaps, i.e. links that either share the same source or the same target word. In a final step, we cluster the binary alignment links using a neighborhood of k source and target words around each link. These clusters now constitute ASR translation phrases. Examples for phrase extracted in this manner, with a neighborhood of $k = 1$, are:

entre Parlamento y Consejo # between parliament and council
presupuesto aqui en el Parlamento # budget here in the parliament

For each of these phrases, we compute the forward and backward translation probability based on the IBM4 lexicons. To incorporate these new phrases into the baseline phrase tables, we extend the baseline phrase table entries with two additional TM probabilities, set to 1 (zero logarithmic cost). Accordingly, the additional ASR phrases have probabilities of 1 at the positions of the four original TM probabilities, followed by the two computed TM probabilities of the ASR phrases. While a visual inspection of the in this manner extracted ASR phrases seemed promising, we only achieved inconclusive results using these phrases. In particular, we observed only a statistically insignificant improvement in one translation direction and a statistically insignificant degradation in the opposite translation direction.

8.3 Biasing Automatic Speech Recognition

The experiments described in the following are based on automatic translations computed with the baseline MT systems.

8.3.1 ASR Language Model Adaptation

Similar to MT language model adaptation, ASR LM adaptation is based on ST n-gram hints, with $n \in \{1, 2, 3\}$. ST n-gram hints are n-grams found in the m-best speech translation hypotheses ($m = 500$) of the ± 6 seconds padded target language audio. Whenever a ST n-gram hint is observed during decoding, a discount is applied to the LM score (cost) of the current ASR hypothesis. In this way, we favor hypotheses that contain ST n-gram hints. The discount value $d(h, n)$ of each ST n-gram hint h depends on the logarithmic baseline LM score of h and is computed as follows:

$$d(h,n) = \begin{cases} w_n * LM_{score}(h) & \text{for } LM_{score}(h) \geq t_n \\ 0 & \text{for } LM_{score}(h) < t_n \end{cases}$$

We estimate optimal values for the parameter w_n and t_n via manual gradient descent on dev05. The threshold t_n ensures that we do not further discounts n-grams with a high LM probability (these are mostly function words). Further, we only apply discounts for ST n-gram hints that can also be found in the 500-best ASR hypotheses of the baseline ASR system. In other words, we only keep those ST n-grams in which baseline ASR and baseline MT agree in their respective m-best hypotheses lists. As ASR decoding is time-consuming, we investigate the effect of the adapted language model not just for ASR decoding, but also for ASR lattice re-scoring. In the following, we refer to applying the adapted LM during decoding as LM_D adaptation and to applying the LM in a re-scoring step after decoding as LM_R adaptation. Accordingly, LM_{R+D} adaptation refers to the combination of both adaptation schemes. Table 8.4 lists the English and Spanish word errors achieved by applying LM_R adaptation to the ASR output of the second pass[1]. We test LM_{R+D} adaptation in an additional third decoding pass in combination with acoustic model adaptation, as described in the following section. For LM adaptation via n-gram discounts, we found that ST uni-gram

[1] As described in detail in Chapter 5, our baseline ASR systems apply a two-pass decoding setup.

		dev05		dtest05	
		En	Sp	En	Sp
2nd pass	baseline	13.2	11.5	9.1	12.7
	LM_R	12.8	11.3	8.7	12.5
3rd pass	baseline	12.7	11.4	8.7	12.4
	AM adapt	12.5	11.3	8.6	12.2
	AM+LM_{R+D}	12.2	11.1	8.4	12.1

Table 8.4: English and Spanish word error rates for biasing ASR with parallel speech in the 2nd and 3rd decoding pass. Biasing schemes include using adapted acoustic models (AM) during decoding and using adapted language models, either in a lattice rescoring step (LM_R) after decoding or both, during decoding and for lattice rescoring (LM_{R+D}).

hints are the most important information source. We observed only very small additional improvements using the bi- and tri-gram ST hints. Similar LM adaptation experiments, reported in Chapter 11, therefore rely only on uni-gram ST hints.

8.3.2 Acoustic Model Adaption

In order to bias the acoustic model with the extracted ST n-gram hints, we add a third decoding pass to our ASR decoding setup. The unsupervised speaker adaptation of the third pass' baseline system relies on the ASR hypotheses of the second pass' baseline system. To bias the AM with the parallel speech in the third pass, we simply rely on hypotheses from the LM_R adapted second pass. We apply word confidences during speaker adaptation in the following manner. Frames associated with words that have a word confidence of $c < 0.5$ are ignored. Frames associated with words that have a word confidence of $c \geq 0.5$ contribute to speaker adaptation with a weight equal to c. The computation of word confidences is strongly influenced by the LM score. For this reason, LM_R adaptation also influences the word confidences computed in the CN combination at the end of the second pass. Speaker adaptation in the third pass is therefore not only influenced by improved ASR hypotheses, but also by changed ASR word

confidences. We achieve lowest word error rates by combining AM adaption with LM_{R+D} adaptation, denoted with AM+LM_{R+D} in Table 8.4.

8.4 Chapter Summary & Discussion

Despite very strong baseline systems and despite strong differences between simultaneous interpretation and translation, we were able to successfully improve ASR and MT system performance by *automatically translating parallel speech (interpretation)* and by then using these translations to adapt the underlying ASR and MT models. The improvements in spoken language translation performance, gained by biasing the involved MT systems and measured in BLEU, are shown in Figure 8.2. The figure compares the BLEU score of the baseline MT system with the BLEU score of the biased MT system. Further, it lists the relatively low BLEU score achieved by scoring the respective target language parallel speech reference transcription against the two target language translation references. These low BLEU scores underscore the strong differences between parallel speech (interpretation) and translation. The reported gains in transcription performance were small, but consistent. We will see in Chapter 11 that parallel speech is of greater value in the context of weaker baseline ASR systems.

Compared to similar experiments that we conducted in the context of biasing dictation systems for human translators [54], the achieved improvements in automatic transcription and translation performance seem small. However, in contrast to our experiments reported in [54], we were exploiting spontaneous parallel speech in the form of EPPS simultaneous interpretations. Our experiments in [54] only considered read translations from the relatively small travel domain.

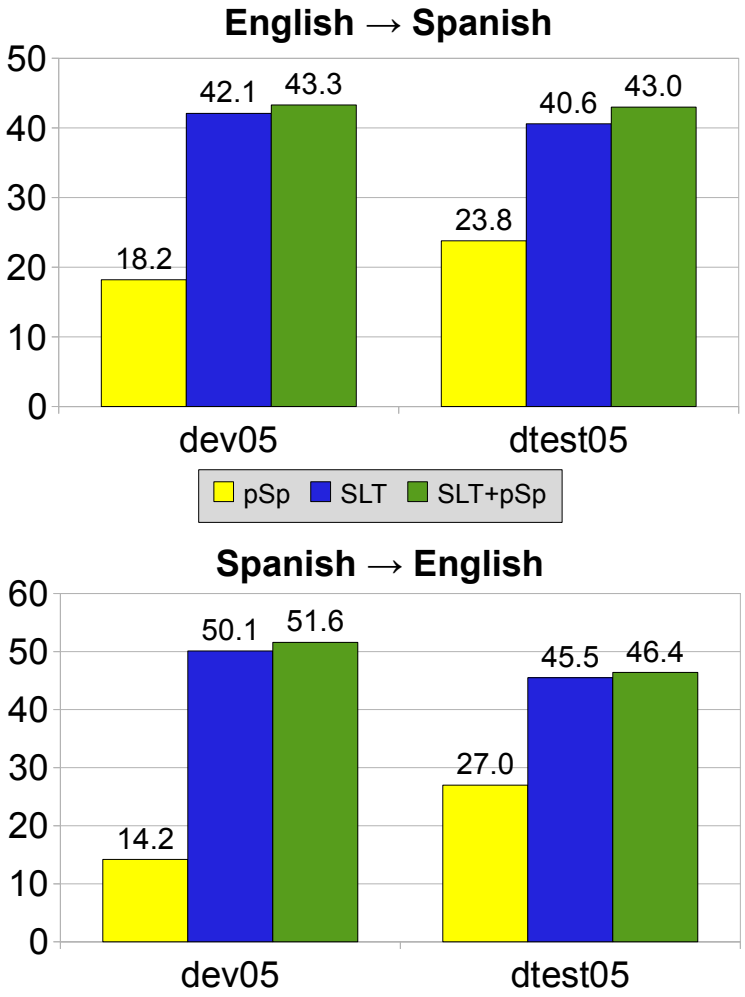

Figure 8.2: Comparing parallel speech (pSp), spoken language translation (SLT) and pSp-biased SLT (SLT+pSp) in terms of BLEU metric.

8. INTERPRETATION AS AN AUXILIARY INFORMATION SOURCE

9

Acoustic Model Training on EPPS Simultaneous Interpretation

9.1 EPPS Data Resource-Limitations

In Chapter 4, we described in detail the enormous translation and interpretation effort that is attached to European Parliament Plenary Sessions (EPPS). Parliamentary proceedings (final text editions, FTEs) are made available in the 23 official languages of the Union, and simultaneous interpretations are provided during the sessions to support the multilingualism of the Members of Parliament. Automatic solutions targeted to support this translation and interpretation effort need to support the many languages spoken in the Parliament. However, large-scale spoken language translation development in the context of EPPS remains mostly limited to the English/Spanish language pair. While large amounts of multilingual parallel text data are available in the form of final text editions, suitable to support the development of translation models and language models, ASR development suffers from an unavailability of—costly—verbatim transcriptions for languages different from English and Spanish. As explained in Section 4.2, the segments of the final text editions that represent transcriptions of the politician speeches are revised versions these speeches and they are formatted for an easy readability. For this reason, FTEs can differ significantly from verbatim

transcriptions of the speeches held in Parliament. Further, the segments of the final text editions that represent translations of the politician speeches deviate even more strongly from their respective simultaneous interpretation. This is not just due to the differences inherent to interpretation and translation, as explained in Chapter 7, but also due to the fact that the FTE translations are based on the already revised and re-formatted FTE transcriptions. Figure 9.1 gives an overview data resources available in the context of European Parliament Plenary Sessions: final text editions and live broadcast audio in the different languages of the Union. The figure also gives an example that compares verbatim transcription of politician speech and interpreter speech with the respective final text edition.

9.2 Lightly Supervised Acoustic Model Training for EPPS

To support the development of EPPS spoken language translation systems for language pairs different from English/Spanish, we consider unsupervised and lightly supervised acoustic model training techniques applied to audio recording harvested from the EPPS live broadcasts [55]. Unsupervised acoustic model training is based on speech data for which no human transcriptions are available. Training relies in that case on automatic transcriptions that are created with an initial ASR system. Lightly supervised acoustic model training [37] refers to the case where some imperfect human transcriptions, for example closed-captions provided during television broadcasts, can be used to either bias the initial ASR system for an improved transcription performance or to filter erroneous ASR hypotheses. Given the availability of final text editions and several audio channels in different languages, recordings of EPPS are suitable for lightly supervised acoustic model training. Supervision for acoustic model training in language L_i can be introduced via the respective final text edition in language L_i or via automatic translations into language L_i from final text editions and interpretations available in languages $L_{j\neq i}$. Gollan et al. mention in [24] the possibility to include English final text editions in the language model training data to achieve a lightly

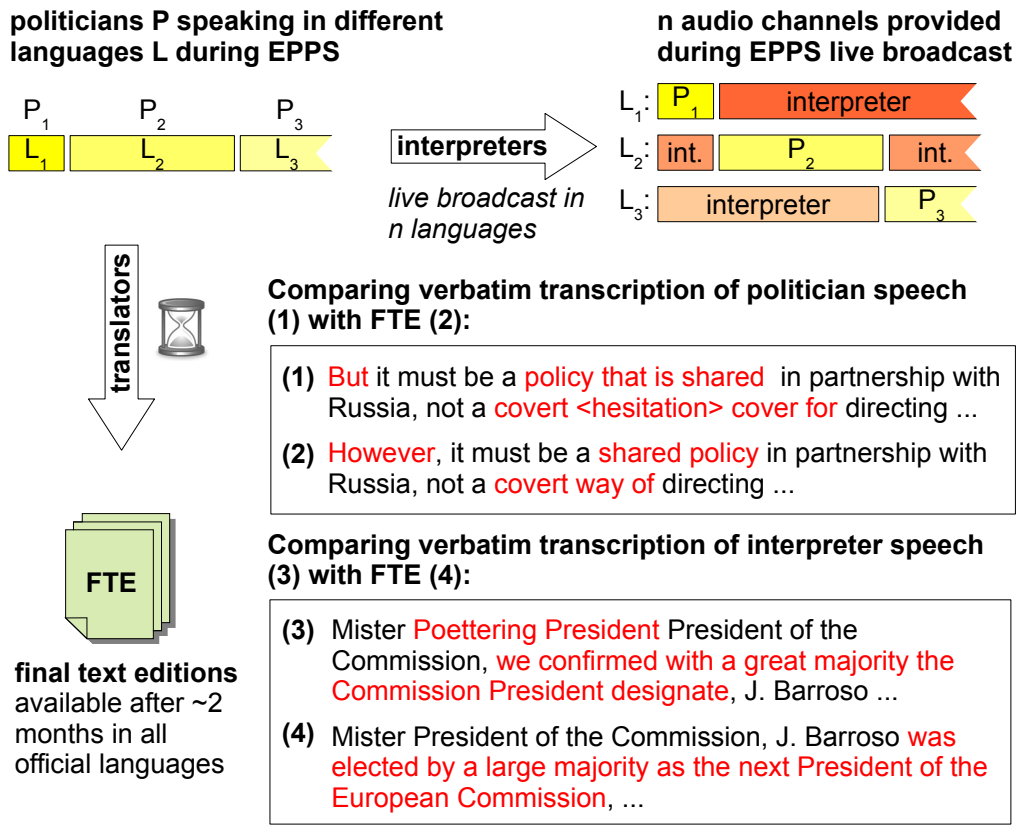

Figure 9.1: EPPS data resources and an example that highlights the differences between final text edition (FTE) and verbatim transcription.

supervised acoustic model training on English EPPS recordings. However, no experiments for FTE-based supervision are reported in [24].

In the following, we examine the impact of FTE-based and pSp-based supervision on German word error rate (WER). We make use of German final text editions and of German automatic translations extracted from the English and Spanish audio channels via spoken language translation. Further, we present results for acoustic model training based on automatic transcriptions created under FTE-based and pSp-based supervision.

9.3 Experimental Setup

9.3.1 Data

Table 9.1 gives an overview on the German audio data statistics of the development set, evaluation set and non-transcribed training set. All sessions included in the respective sets were recorded in our laboratory from the EPPS satellite live broadcasts in several languages, including German, English and Spanish. For audio segmentation, we use the same language-independent Hidden Markov Model based speech/non-speech audio segmenter as we applied it on the English/Spanish dev06 and eval07 sets. The speakers in the recordings switch between interpreters and native or non-native speakers with different, and partly strongly pronounced, accents. Whenever a politician speaks in a language different from the language of the respective audio channel, the microphone is switched back to the interpreter. Delays when the microphone is switched result in short periods of foreign language speech. The German verbatim transcriptions used to measure automatic transcription performance were created in our laboratory. To measure automatic translation performance, we compute BLEU scores based on these German verbatim transcriptions. We did not we create any additional English→German or Spanish→German reference translations. Therefore, the presented BLEU scores actually measure the similarity of German automatic translations, created from English or Spanish parallel speech, to German parallel speech. We do not have any reference transcriptions available for the English and Spanish parallel speech. Therefore, it is not possible to compute the English and Spanish WER.

	sessions	utterances	audio [h]
dev	1	592	1.69
eval	1	885	2.04
training	93	73,408	142.74

Table 9.1: German audio data statistics: development, evaluation and training sets.

9.3.2 ASR Systems

The **German ASR** system features a speaker-independent and a speaker-dependent decoding pass. The ASR subsystems used in the two passes feature MVDR front-ends [81]. In the first pass a single ASR system is employed to provide the second pass with first-best hypotheses for unsupervised speaker adaptation. We apply Maximum Likelihood Linear Regression, feature space constrained MLLR and vocal tract length normalization for speaker adaptation. The second pass uses two ASR subsystems with slightly different phones sets. At the end of the second pass, confusion network combination [40] is applied. The acoustic models are trained on 70h of German broadcast news data[1]. The dictionary consists of 89.6k pronunciation entries. Compound word splitting is employed to keep the out-of-vocabulary rate small. The 4-gram LM is trained on the German final text editions extracted from the Europarl v3 corpus [32]. The LM perplexity and WER for the development set and evaluation set are shown in the first column of Table 9.4.

The **English ASR** and **Spanish ASR** are identical to the EPPS English and Spanish ASR systems described in Chapter 5. The English and Spanish WER on the used development set and evaluation set cannot be computed, since we do not have the necessary English and Spanish verbatim transcriptions available. The typical word error rate on this task ranges from 11% to 12% for the Spanish ASR system and from 12% to 13% for the English ASR system.

9.3.3 MT Systems

The **English→German** and **Spanish→German** MT systems are trained on the respective parallel parts of the Europarl v3 corpus [32]. Training data statistics are shown in Table 9.2 and Table 9.3. We lowercased the training data and removed all punctuation. Both systems apply the same 4-gram language model as used in the German ASR system. For decoding, we rely on the internal word reordering of the ISL decoder. We use a reordering window of 2.

[1]Thanks go to Christian Fügen and Matthias Wölfel for providing the baseline acoustic models and to Florian Kraft for providing the German compound word splitter.

	English	German
sentence pairs	1212846	
unique sent. pairs	1194175	
sentence length	25.4	23.4
words/tokens	30.8 M	28.4 M
vocabulary	85.8 K	284.6 K

Table 9.2: MT training corpus statistics, English→German.

	Spanish	German
sentence pairs	1257807	
unique sent. pairs	1238129	
sentence length	26.5	24.2
words/tokens	33.3 M	30.4 M
vocabulary	128.2 K	189.7 K

Table 9.3: MT training corpus statistics, Spanish→German.

The English→German and Spanish→German automatic translations are computed using the first-best English/Spanish ASR hypotheses as input. The translation reference for IBM BLEU score computation is equal to the German reference transcription used for computing the German ASR word error rate. The BLEU score on the development set is 12.5 for English→German and 11.9 for Spanish→German. On the test set, the BLEU score is 15.2 and 13.4, respectively. Although the Spanish ASR word error rate is typically approximately 1% lower than the WER of the English ASR, we see a consistently better translation performance for English→German translation. This can be explained by the fact that Spanish is a morphologically rich language compared to English.

9.4 FTE-based and pSp-based Supervision: Impact on WER

We employ two different types of supervision that are based on the final text editions. In the case where the FTE of a specific session is part of the overall

language model training data, we speak of a 'general' FTE (gFTE) supervision. Whenever we mention FTE supervision without the addition of the word 'general' we refer to the case where the final text edition of a specific session receives a higher weight than the remaining LM training data. We achieve FTE supervision by building a language model on all available FTEs and a LM on the FTE of the respective session. We then interpolate both language models with a fixed interpolation weight of 0.28 for the smaller, session specific language model. This interpolation weight was determined to yield the lowest perplexity on the development set.

In order to introduce pSp-based supervision, we automatically transcribe and translate the English and Spanish audio that is available for each session. Using the 1000-best translation hypotheses from each MT system, we build two separate language models and interpolate these. The LM based on the English→German translations receives an interpolation weight of 0.54. Finally, we interpolate this LM with the FTE supervised LM, where the interpolation weight for the translation based LM is set to 0.34. The used interpolation weights are again determined on the development set to yield a minimal perplexity.

The English and Spanish ASR systems do not employ any form of supervision, that is, we do not apply FTE supervision nor pSp supervision to these systems. However, our English→German and Spanish→German MT systems apply general FTE supervision on development, test and training set, since the respective final text editions are part of the translation model and language model training data.

Table 9.4 shows the German language model perplexity and German word error rate for the different types of supervision. The first column gives the results when no supervision of any form is applied. In the last column we list the results for combining FTE supervision with pSp-based supervision using both, English and Spanish parallel speech (interpretation), as well as the results when using only the English or Spanish pSp on top of FTE based supervision. The results show that a significant gain in transcription performance can be achieved by applying

supervision→		none	gFTE	FTE	FTE & pSp
PPL	dev	219	206	161	138_e 142_s **130**
	eval	190	176	146	127_e 130_s **118**
WER	dev	22.3	21.6	20.9	20.7_e 20.3_s **20.1**
	eval	21.0	20.1	19.4	19.1_e 19.2_s **18.8**

Table 9.4: Language model perplexity (PPL) and word error rate (WER) for different types of supervision. The last column lists results for combining FTE supervision with pSp-based supervision using either English parallel speech ($_e$), Spanish parallel speech ($_s$) or both, English and Spanish parallel speech together.

FTE supervision and a combination of FTE supervision and pSp-based supervision. Further, the results suggest that the gain in transcription performance for pSp-based supervision depends on the number of languages used. This indicates that complementary information is added with each additional language.

9.5 FTE and pSp Supervised Acoustic Model Training

We automatically transcribe the available 142.7h of German EPPS recordings, using general FTE supervision, session specific FTE supervision and a combination of session specific FTE and pSp supervision. Before training on these automatic transcriptions, we apply a simple rule based filter to remove noisy and low confidence utterances. The rules for this filter are hand written and tuned on the development set in regards to word error rate. We remove all utterances that have a filler to word ratio that is greater than 2.5 or that have an average word confidence lower than 0.4. During training we do not apply the available word confidence scores in any way. Acoustic model training consists of two iterations of Viterbi training starting from the baseline AM.

To test the new acoustic models, we add a simplified third decoding pass to the German decoding setup. This simplified pass features only one ASR system and does not include confusion network combination. Unsupervised speaker

adaptation is performed on the confusion network combination output from the second pass. Results are listed in Table 9.5. The word error rate for the original acoustic model in the third pass is slightly worse than the WER achieved after confusion network combination in the second pass. We list results achieved with the different AMs using either no supervision on the development and evaluation set or a combination of FTE-based and pSp-based supervision. Compared to the original AM, the re-trained AMs show in both cases significant gains in transcription performance. Using the best-performing AM (FTE & pSp AM) and applying no supervision on the evaluation set, the 21.0% WER of the baseline system is reduced to 19.8%. Applying FTE+pSp supervision on the eval set, the WER is further reduced to 18.0%.

supervision→	**none**		**FTE & pSp**	
	dev	eval	dev	eval
original AM	22.2	21.2	20.6	19.2
gFTE AM	20.9	20.0	18.8	18.4
FTE AM	20.7	19.9	18.8	18.2
FTE & pSP AM	20.4	19.8	18.8	18.0

Table 9.5: Word error rates achieved in a third decoding pass, using different acoustic models (AM) and applying either no supervision or FTE & pSP based supervision.

9.6 Chapter Summary & Discussion

We achieved significant gains in German transcription performance by applying unsupervised and lightly supervised AM training in the context of audio recordings harvested from EPPS live broadcasts. Thus, we successfully exploited the live broadcast EPPS simultaneous interpretations and speeches as acoustic model training data without having to create costly verbatim transcriptions. In combination with the language dedicated audio channels provided during EPPS, the proposed approach therefore supports the development of ASR systems in the various languages of the European Union. Our approach introduces light supervision on a per session basis by using the freely available EPPS data resources:

final text editions and live broadcast parallel speech. Light supervision was either applied during training only or during training and testing. However, while the applied light supervision techniques resulted in significantly lower word error rates—on the evaluation set, the word error rate between gFTE and FTE & pSp supervision differs by 1.3% absolute in the second pass—the impact of these lower word error rates during training seems to remain small. For example, in the third pass and applying no supervision on the evaluation set, the gFTE acoustic model is only 0.2% absolute worse than the FTE & pSp acoustic model.

10

Automatic Translation from Simultaneous Interpretation

In this chapter we introduce our approach for training statistical translation models from parallel speech audio [56]. Our experiments are based on parallel speech audio from English/Spanish simultaneous interpretation, as provided during EPPS. As motivated in the introduction, training translation models from parallel speech audio is of special interest for speech translation development between new language pairs where pre-existing data resources for traditional speech translation training are scarce. The significant data resources that are already available in the context of EPPS enable us to study our approach at different levels of resource availability.

10.1 General Approach & Major Challenges

We use parallel speech audio within a standard training setup for phrase-based statistical machine translation. To do so, we transcribe the parallel speech using ASR. The resulting ASR hypotheses are aligned on a speech utterance basis by applying special alignment strategies that are tailored to the parallel speech of simultaneous interpretation. These alignment strategies are described in detail in Section 10.3. With the help of these strategies, we align to each automatically transcribed source speech utterance the related target speech transcription. This forms the first part of our bilingual TM training corpus. The second part results

from repeating the same utterance alignment procedure in the reverse direction. Our standard training setup extracts phrase tables from the created bilingual training corpus by using the GIZA++ toolkit [49], in combination with University Edinburgh's training scripts, as provided during the NAACL 2006 Workshop on Statistical Machine Translation [33]. The GIZA++ toolkit is run with its standard parameter settings.

The major challenges faced include (a) the significant difference between translation and interpretation as explained in Chapter 7, (b) the problem of aligning source and target speech utterances and (c) the (high amount of) transcription errors in the ASR output (of the resource-deficient ASR).

10.2 Experimental Setup

10.2.1 Data and Scoring

For training translation models from parallel speech, we use a parallel speech corpus that was recorded in our laboratory from satellite live broadcasts of EPPS. These recordings do not include any parliamentary sessions used for ASR system training nor do they overlap with our development or evaluation sets. For segmenting the pSp corpus, we use the same language-independent Hidden Markov Model based speech/non-speech audio segmentation as we applied it on dev06 and eval07. To measure the performance of the trained translation models, we use the English/Spanish dev06 and eval07 sets, described in detail in Section 4.4. As dev06 and eval07 only consist of politician speech and do not form a pSp corpus, we further rely on dev05 to tune our alignment algorithm presented in Section 10.3. Table 10.1 summarizes the data statistics of the used pSp corpus. The amount of words included in the parallel speech corpus is estimated on the ASR first-best hypotheses, since no reference transcriptions are available for the pSp corpus.

For scoring ASR and MT performance we use non-punctuated, lowercased references. MT performance is measured in IBM BLEU. We use the mWER

	English	Spanish
utts [k]	65.3	63.2
words [k]	954.4	897.0
audio [h]	111.3	105.2

Table 10.1: Parallel speech corpus: amount of utterances, words and audio.

segmentation script, to align the translated speech utterances to the translation references for scoring.

10.2.2 ASR and MT Systems

English and (unconstrained) Spanish ASR are identical to the systems described in Chapter 5. In addition to the standard Spanish ASR system we use two constrained Spanish ASR systems, Sp_{c0} and Sp_{c1}, to simulate ASR performance levels encountered in the context of resource deficiency. In the situation of resource limitation, the lack of text data and transcribed audio data leads to a weak LM and and a weak AM. Both contribute to an increased WER. To simulate resource limitation, we first (Sp_{c0}) constrain the Spanish LM to the 748k running words of the transcriptions that were used to train the AM. To simulate a lower performance AM (Sp_{c1}), we further limit the system to a context independent phone-set. That is, system Sp_{c1} applies the constrained LM in addition to a constrained AM. Table 10.2 lists the word error rates and LM perplexities of the used ASR systems.

	dev06				**eval07**			
	Sp	Sp_{c0}	Sp_{c1}	En	Sp	Sp_{c0}	Sp_{c1}	En
PPL	89	178	178	108	89	177	177	106
WER	8.4	16.1	33.3	13.9	9.0	16.5	33.1	12.2

Table 10.2: Language model perplexity (PPL) and ASR word error rates (WER).

For MT, we use the ISL beam search decoder. The reordering window of the internal word reordering model is set to 4. Spoken language translation is achieved by applying the MT system on the ASR first-best hypotheses. The MT

language models are identical to the language models used in the ASR systems. As a consequence, English→Spanish spoken language translation relies in the context of a simulated resource-limitation on the constrained Spanish LM.

10.3 Aligning Parallel Speech of Simultaneous Interpretation

Where not otherwise mentioned, the pSp corpus used in this section is transcribed at estimated English and Spanish WER levels of 12-14% and 9%, respectively. We estimate these numbers based on the English and Spanish ASR performance on dev06 and eval07, since no manual transcription of the pSp corpus is available.

Since simultaneous interpreters have to keep pace with the source language speaker, an approximate time alignment between source and target language speech is already given. We can exploit this fact to align source speech utterances to parallel target speech by considering the target speech snippet that starts/ends x seconds before/after the source speech utterance starts/ends. We need to include target speech before the start time of the respective source utterance since we do not know which of the audio channels contains interpreter speech. In fact, it often occurs that both audio channels contain interpreter speech. In such cases, the politician that took the floor in the Parliament is giving a speech in a language different from English and Spanish. To minimize computation time, we decode the pSp corpus only once, based on the speech utterance segmentation that was introduced via voice activity detection prior to ASR. To extract the ASR hypotheses of the padded speech snippets, we rely on the hypothesized word-start and word-end times.

In order to find an optimal padding value x, we conduct two sets of experiments. First, on dev05 and for different values of x, we compute the F1-measure in respect to uni-gram matches between the padded, automatically transcribed pSp snippets and the dev05 available reference translations. Figure 10.1 depicts how the F1-measure changes for different values of x. It shows a peak at $x = 2$

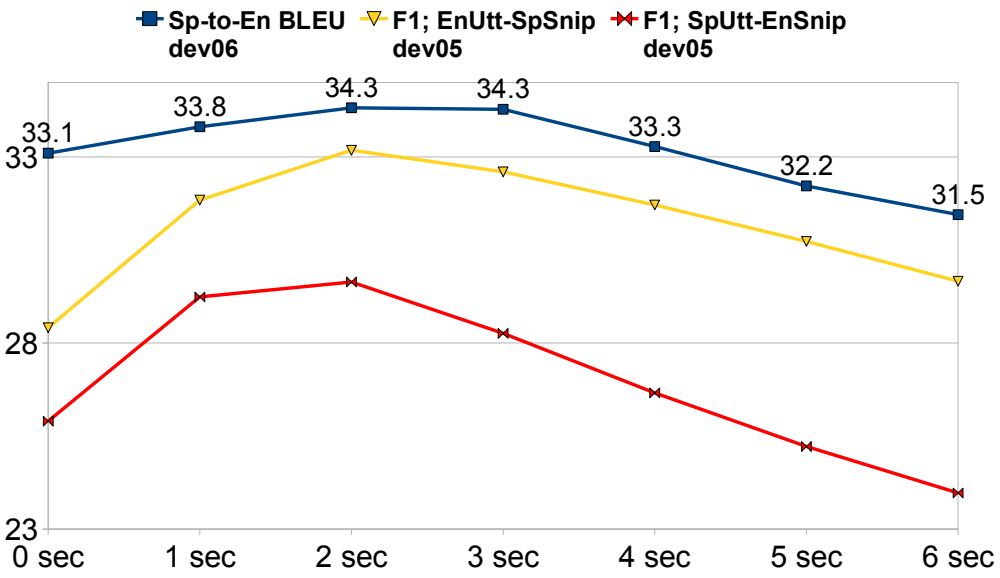

Figure 10.1: F1-measure (y-axis) on dev05 and BLEU score on dev06 for different target speech snippet padding values $x \in \{0, 1, ..., 5, 6\}$.

seconds for both cases, when aligning English utterances to Spanish pSp snippets and when aligning Spanish utterances to English pSp snippets. In the second set of experiments, we create seven different parallel MT training corpora from the automatically transcribed pSp; one training corpus each for $x \in \{0, 1, ..., 5, 6\}$ seconds. After extracting seven different phrase tables from these MT training corpora, we compute the translation performance for Sp→En on dev06, using these phrase tables. As we can see in Figure 10.1, the BLEU score again peaks at $x = 2$s, showing that the F1-measure computed on dev05 correlates well with the translation performance on dev06. In other words, the optimal padding value x for aligning our pSp corpus can be well predicted by simply computing the F1-measure on dev05.

In addition to a simple word-time based padding of the parallel speech snippets for aligning the pSp corpus, we also experiment with a more sophisticated two-pass alignment strategy. By manually inspecting the parallel speech present in dev05, we find that, if the information contained in the source utterance is at all present in the parallel speech, a 6 seconds utterance padding almost al-

ways guarantees that the information can be found in the respective target audio snippet. By a 6 seconds utterance padding, we refer to the case where a target speech snippet is comprised of all target speech utterances that fall into the time window that is formed by padding the source utterance start/end time with 6 seconds. Figure 10.2 gives an example of parallel speech that is aligned based on a 6 seconds utterance padding. In addition to the transcription reference of the Spanish speech utterance and the respective English pSp-snippet, the figure shows one of the two Sp→En translation references. The part of the English speech snippet that is directly related to the Spanish speech utterance is in red font. As can be seen, the padded pSp segment contains too much irrelevant, and potentially misleading, information.

Our two-pass algorithm for aligning parallel target speech to source speech utterances operates on a per-source-utterance basis and uses 6 second utterance-padded target speech snippets. In addition to the source utterance at hand, the algorithm also considers all neighboring source utterances that overlap in their respective target speech snippet with the target speech snippet of the current source utterance. In a first step, the combined forward and backward translation probability for each source word to each target word is computed and an alignment link is introduced if the combined translation probability is above a specific threshold t_p and if the absolute distance between source word-start time and target word-start time is below a specific threshold t_d. The word-to-word translation probabilities are based on IBM4 word lexicons that are computed in a first pass on the parallel MT training corpus that results from a 2 second word-time based padding of the pSp snippets. The translation probability of the alignment link al between source word sw and target word tw is weighted by the combined translation probability times the 'importance' of the target word. We define the importance of the target word as:

$$importance(tw) = 1.0 - sL * LM(tw) \tag{10.1}$$

with sL equal to the length in words of the target speech snippet and $LM(tw)$ equal to the uni-gram LM probability of the target word. In a next step, we find

SPANISH UTTERANCE: "hay que terminar los las negociaciones"

TRANSLATION: "**it is necessary to complete the negotiations**"

PARALLEL SPEECH SNIPPET: "so we have our buildings policy for brussels there are two buildings in particular that account for more than three hundred million euros the two buildings there in brussels **we have negotiations under way they shall be concluded** i hope before the end of the year and that would mean that we could already start making some of the payments in the year two thousand and five"

SPANISH UTTERANCE: "el las conversaciones para que ya podamos hacer una parte de esos pagos en el año dos mil cinco"

TRANSLATION: "**and the conversations to enable us to make a part of those payments already in the year two thousand five**"

PARALLEL SPEECH SNIPPET: "buildings in particular that account for more than three hundred million euros the two buildings there in brussels we have negotiations under way they shall be concluded i hope before the end of the year and **that would mean that we could already start making some of the payments in the year two thousand and five** also in addition to that we are going to try to make sure that members of staff from different members states of the european union will be granted an equal status"

SPANISH UTTERANCE: "también"

VERBATIM TRANSLATION: "**in addition**"

PARALLEL SPEECH SNIPPET: "and that would mean that we could already start making some of the payments in the year two thousand and five also **in addition** to that we are going to try to make sure that members of staff from different members states of the european union will be granted an equal status"

SPANISH UTTERANCE: "trataremos de que todo el personal tenga"

TRANSLATION: "**we shall try that all the staff will get**"

PARALLEL SPEECH TRANSCRIPT: "in addition to that **we are going to try to make sure that members of staff** from different members states of the european union **will be granted** an equal status we look forward to amend the statuate of course we hope that that will be approved as soon as possible and we hope that it proves viable in practice"

Figure 10.2: Examples for pSp based on a 6 seconds utterance based padding.

the optimal 'left cut' position in the target speech snippet that defines all words before this position as irrelevant to the source utterance and all words after this position as relevant. This is done by computing the sum over all alignment links left of a cut position candidate that belong to neighboring source utterances and then adding the sum computed over all alignment links right of the cut position candidate that belong to the current source utterance. The cut position with the highest overall sum is selected. During this process, we also consider alignment link clusters forming target bi- and tri-grams. For each such cluster we introduce additional alignment links that are included in the overall sum. The alignment link $alBI$ for a bi-gram alignment cluster formed by the alignment links al_1 and al_2 is, for example, given as:

$$alBI(al_1, al_2) = (al_1 * al_2)^{bw} \tag{10.2}$$

with the bi-gram weight bw to allow for a flexible additional weighting of such bi-gram link clusters. Accordingly, an optimal right cut position is found by computing the sum over all alignment links left of the cut position that belong to the current source utterance and adding the sum computed over all alignment links right of the cut position that belong to neighboring source utterances.

To optimize the two-pass alignment algorithm, we perform a grid search on dev05, aiming for a maximal value of F1-measure that is based on matching uni-grams in the pSp snippets and the reference translations. In addition to uni-gram F1-measure (and precision and recall), we also compute the respective values for n-gram matches with $n \in [1, 4]$. Table 10.3 shows the results for the two alignment passes of the algorithm. The first pass is identical to the 2 seconds word-time based padding of the speech snippets. The table shows that the two-pass algorithm yields higher F1 values at a higher precision and lower recall than the word time based padding. Further, we can see that the overall low recall degrades strongly for higher order n-grams. We observe values as low as 3.4 for 4-gram matches between the parallel speech snippets and the two reference translations. This underlines the strong difference between translation and interpretation, as

already explained in detail in Chapter 7.

Table 10.4 lists the Spanish→English machine translation performance when using the two different alignment strategies and automatically transcribing the parallel speech corpus at different Spanish word error rate levels. At all three Spanish word error rate levels, the two-pass alignment strategy outperforms the word-time based alignment by approximately 1 BLEU point. This is in all cases statistically significant ($p < 0.05$). The results also show that, even for a highly degraded Spanish transcription performance at 33% WER (3.7 times worse than the transcription performance of the standard Spanish ASR system), the machine translation performance degrades only by approximately 12% relative on the eval set. This indicates that training translation models from automatically transcribed parallel speech is robust to strong variations in ASR performance on one side of the parallel speech corpus.

n	**alignment**	**EnUtt-SpSnip**			**SpUtt-EnSnip**		
		Pre	Rec	F1	Pre	Rec	F1
1	2 seconds	34.8	31.7	33.2	24.9	36.7	29.6
	2-pass	38.9	30.8	34.4	29.7	35.0	32.1
2	2 seconds	15.2	12.4	13.7	9.6	13.6	11.2
	2-pass	17.5	12.7	14.7	11.6	13.9	12.7
3	2 seconds	8.2	6.4	7.1	4.8	6.8	5.6
	2-pass	9.7	6.6	7.9	5.7	6.9	6.3
4	2 seconds	4.6	3.4	3.9	2.5	3.7	3.0
	2-pass	5.6	3.7	4.5	3.0	3.8	3.3

Table 10.3: Precision, Recall and F-measure (F1) on dev05 for the two utterance alignment passes.

	dev06			eval07		
Sp pSp WER	9%	16%	33%	9%	16%	33%
2 seconds	34.3	32.1	28.2	33.5	32.6	28.5
2-pass	35.1	33.5	29.1	34.3	33.5	30.1

Table 10.4: 2-pass alignment strategy: Sp→En automatic translation performance using Spanish reference transcriptions (0% word error rate) as input to MT and translation models trained with parallel speech transcribed at Spanish word error rate levels of 9/16/33%.

10.4 Machine Translation and Speech Translation Results

Table 10.5 lists the Sp→En and En→Sp machine translation results obtained when using translation models trained from parallel speech. We list the results for parallel speech that was transcribed at different Spanish word error rate levels; the approximate Spanish WER achieved on pSp is shown in the first column. The English ASR system was kept unchanged; we estimate its WER at approximately 12-14% WER, given its performance on dev06 and eval07. BLEU scores marked with $_c$ were computed using the constrained Spanish LM. For comparison, we also list results when training translation models on a bilingual, sentence-aligned text corpus of manual translations. This text corpus was extracted from the bilingual MT training corpus as it was provided during the TC-STAR evaluation. We randomly selected sentence pairs from the original TC-STAR training corpus, until the number of running words on the English part reached 954.4k running words. This is the same number of running words as we estimated for the English part of our pSp corpus. The TC-STAR training corpus is based on the final text editions. It therefore exhibits a certain mismatch in style compared to verbatim style transcriptions and translations. To reduce this mismatch we pre-processed the text corpus accordingly. This pre-processing included punctuation removal, expansion of abbreviations and conversion of numbers and dates to their spoken form.

training corp.	**dev06**		**eval07**	
type, WER	Sp→En	En→Sp	Sp→En	En→Sp
translations, 0%	44.5	48.0	43.9	40.9
pSp, ∼ 9%	35.1	39.7	34.3	31.2
pSp, ∼16%	33.5	34.5_c	33.5	27.0_c
pSp, ∼33%	29.1	31.1_c	30.1	23.3_c

Table 10.5: Training corpus dependent MT performance in BLEU. Results are shown for using a translation model training corpus of manual translations (first data row) or a training corpus of parallel speech (pSp). The pSp corpus was automatically transcribed at three different Spanish word error rate levels (9/16/33%).

The results show degraded translation performance for training translation models from pSp, compared to using a bilingual text corpus of manual translations for training. Using our best-performing ASR systems, the absolute degradation amounts to approximately 10 BLEU points for both translation directions. This degradation in performance results from (a) word errors introduced by automatically transcribing English and Spanish speech (b) the mismatch between translation and interpretation, and (c) errors when aligning the interpreter speech. Nevertheless, we are able to report surprisingly high BLEU scores of up to 34.3 for Sp→En at WER levels of approximately 9% for Spanish ASR and 12-14% for English ASR. As already noted at the end of Section 10.3, we observe only a relatively small degradation in MT performance when introducing a strong degradation in Spanish ASR performance from approximately 9% to 33% WER.

In Section 7.3, we estimated the 'translation' quality, in terms of BLEU, of dev05 parallel speech by comparing the manual transcription of this parallel speech with the two reference translations available for dev05. The BLEU scores were 18.2 when comparing English pSp with the two Spanish→English reference translations, and 14.4 when comparing Spanish pSp with the two English→Spanish reference translations. Our best pSp-trained translation models achieve BLEU scores of 43.5 and 34.8 for Spanish→English and English→Spanish, respectively. These scores are 2 to 3 times higher than the scores of the pSp given in dev05.

Therefore, our pSp-trained translation models, while being trained on parallel speech, are able to achieve automatic translations that are significantly more similar to manual translations than parallel speech itself.

The highest achieved Sp→En translation performance of 34.3 BLEU is on the same level as the translation performance of a translation model trained on 100k English words of sentence aligned translations. We approximate the number of English words in the pSp corpus to be 954.4k. This suggests that, at the considered WER level, translation models trained on pSp audio of En/Sp simultaneous interpretation require 10 times more data (measured in number of translated/interpreted words) than translation models trained on manual translations, to reach similar BLEU performance. Figure 10.3 depicts the development of BLEU score depending on a successively increased training corpus size in 100k word increments, using either a training corpus of translations or our pSp corpus transcribed at a Spanish WER of 16%. The absolute difference between the BLEU scores of both types of translation models is higher for smaller training corpus sizes. At a corpus size of 100k English words, the difference is 15.7 points (a 46.2% relative degradation) and levels out at 500k words to approximately 10.5 to 11 points (a 24.0% to 26.4% relative degradation). Further, we observe that the corpus-size dependent development of BLEU score of the pSp-trained TM mirrors the development of BLEU score seen for the traditionally trained TM, just at a lower level.

Table 10.6 lists the speech translation results on eval. The word error rate on the respective eval source text is shown in the second row. BLEU scores marked with $_c$ were achieved using the constrained Spanish LM. We used the same decoder weights found via minimum error rate optimization on the dev06 verbatim transcriptions, as we had good experience in the past with this approach on the very same development and test sets. For this reason, we do not provide speech translation results for dev06. Compared to translation models trained on a similarly sized bilingual text corpus of translations, we observe a degradation of approximately 8 BLEU points when using parallel speech trained translation models. This degradation in performance is almost 2 BLEU points less than

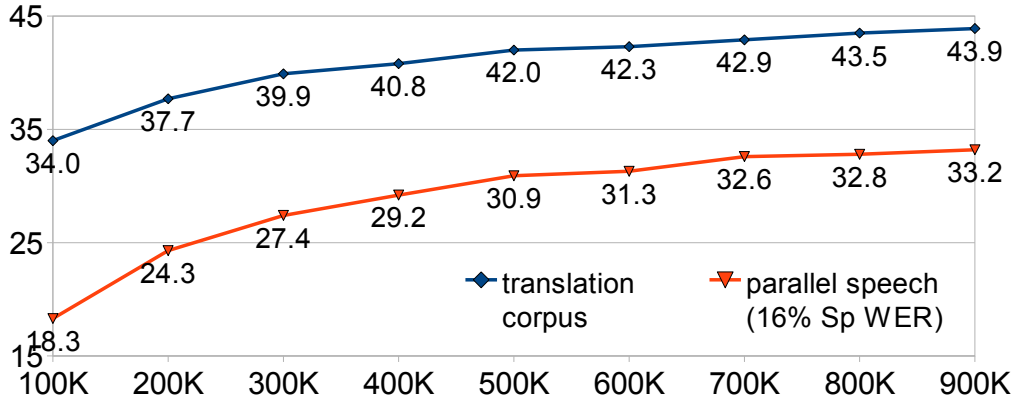

Figure 10.3: BLEU score (y-axis) dependent on training corpus type (parallel speech vs. manual translation) and training corpus size in steps of 100k running words.

in the case of MT (compare Table 10.5). In general, the relative degradation in BLEU caused by recognition errors in the input text is smaller for parallel speech trained translation models (compare Figure 10.4). For example, a WER of 16.5% in the Spanish input to the translation system causes the BLEU score to degrade by 18%, from 43.9 to 36.0, if the system is trained on manual translations. However, the BLEU score of the system trained on pSp audio degrades only by 13.1%, from 33.5 to 29.1. This smaller relative degradation due to word errors in the source input can be observed for all three investigated Spanish WER levels. We apply the same ASR systems used for transcribing the pSp corpus when we automatically transcribe the source speech of the evaluation set for speech translation. The smaller degradation in BLEU score indicates that the parallel speech trained translation models are able to compensate for word errors in the source ASR by incorporating mappings between source word errors and correct target translation. This ability to compensate for source word errors helps to attenuate the loss in speech translation performance experienced by using parallel speech of simultaneous interpretation instead of manual translation for translation model training.

training corp.		**Sp→En**			**En→Sp**
type, WER		**9.0%**	**16.5%**	**33.1%**	**12.2%**
translations, 0%		40.0	36.0	27.8	33.8
pSp., ∼ 9%		31.5	-	-	26.1
pSp., ∼16%		-	29.1	-	22.8_c
pSp., ∼33%		-	-	21.0	19.8_c

Table 10.6: Training corpus dependent ST performance in BLEU. The word error rate of the ASR first-best hypotheses used as machine translation input are shown in bold font. Translation model training is either based on a training corpus of manual translations (first data row) or on a training corpus of automatically transcribed parallel speech (pSp). The pSp corpus was automatically transcribed at three different Spanish word error rate levels (9/16/33%).

10.5 Chapter Summary & Discussion

We created a MT training corpus from non-transcribed parallel speech of simultaneous interpreters by automatically transcribing and aligning source language and target language speech. This enabled us to build MT systems and speech translation systems from simultaneous interpretation, thus eliminating the need for a manually created text corpus of sentenced aligned translations. Our experiments show that in the case of speech translation, parallel speech trained translation models profit from an ability to compensate for word errors in the source ASR. Further, we have shown that training translation models from parallel speech is robust towards low transcription performance on one side of the automatically transcribed speech corpus.

We achieve surprisingly strong translation results with our parallel speech trained translation models. Based on these results, we argue that interpreter speech can present a valuable resource for training MT and speech translation in the context of resource-deficient languages. However, our experiments remain limited to simultaneous interpretation. We believe that the prevailing form of interpretation in the context of resource-deficient languages is consecutive interpretation, as simultaneous interpretation typically demands considerable amounts

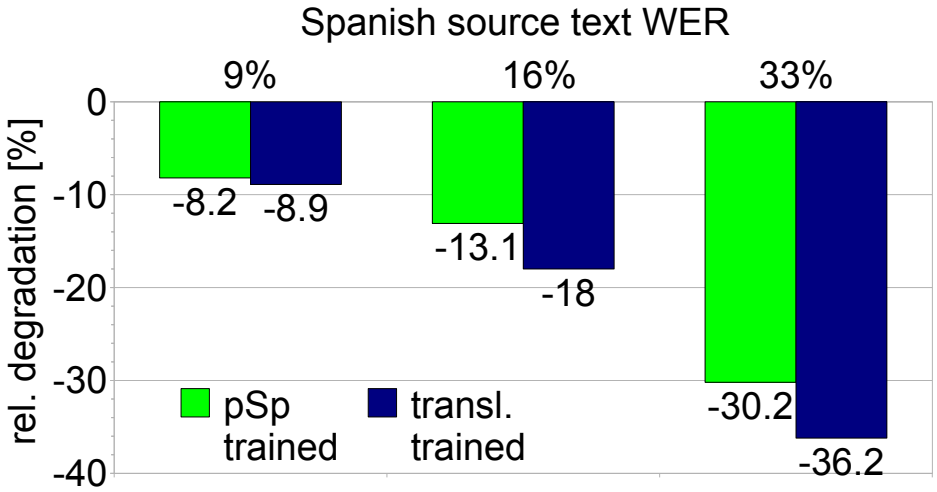

Figure 10.4: Relative degradation in BLEU (Sp→En), dependent on Spanish input word error rate and training corpus type (parallel speech, pSp vs. manual translations).

of expensive equipment (sound proof booths, etc.). For this reason, we investigate the training of translation models from the parallel speech of consecutive interpretation in Chapter 12.

10. AUTOMATIC TRANSLATION FROM SIMULTANEOUS INTERPRETATION

11

Interpretation as Speech Translation Training Data

In the previous chapter, we demonstrated that statistical translation models can be trained in a fully automatic manner from audio recordings of simultaneous interpretations. In this chapter, we extend the use of parallel speech audio as a data resource for unsupervised and lightly supervised training of all major models involved in statistical speech translation: the ASR acoustic model and ASR language model as well as the MT translation model and MT target language model. Specifically, we explore techniques for training acoustic models, language models and translation models from automatically transcribed parallel speech [57]. The parallel nature of pSp audio does not only allow us to train translation models, as described in detail in the previous chapter, but it also allows us to introduce light supervision for model training, as explained in detail in this chapter. We conduct our experiments on a subset of the English/Spanish parallel speech corpus from Chapter 10. To simulate the setting of speech translation between a resource rich and a resource-deficient language, we limit the supervised training data for the Spanish models to 10h of manually transcribed Spanish audio and to a parallel text corpus of sentence-aligned, manual translations that comprises 100k of Spanish words translated into English. Similar to previous experiments, we also consider the situation where only parallel speech audio and no parallel text data of sentence-aligned manual translations is available.

11.1 System Architecture

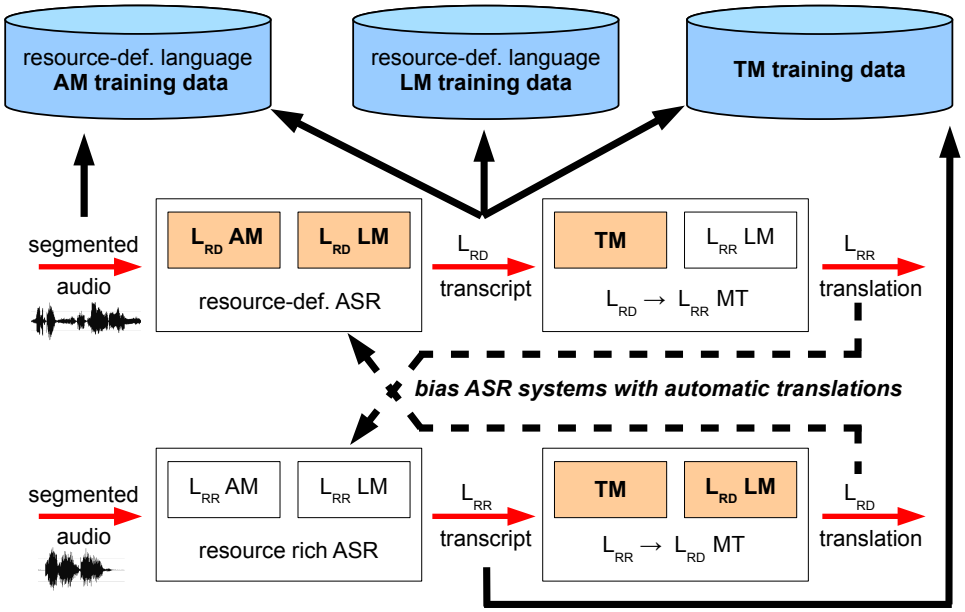

Figure 11.1: Extracting speech translation training data from parallel speech.

In the proposed scenario of speech translation between a resource rich language L_{RR} and a resource-deficient language L_{RD}, we seek to improve the statistical ST models that suffer from the resource deficiency, by automatically creating training data from pSp audio. Figure 11.1 shows our system architecture. The overall system consists of two ST sub-systems, each featuring an ASR component and a MT component. The ASR systems accept pre-segmented speech utterances; we use a HMM-based, language-independent speech/non-speech audio segmentation. The models affected by the resource deficiency are highlighted in color in the diagram. The core components necessary to create ST training data from pSp audio are the two ASR systems. Together with the input audio, automatic transcriptions for L_{RD} can be used for unsupervised AM training. The transcriptions can also be used as additional LM training data. Further, the hypotheses of both ASR systems can be tied together in a parallel training corpus suitable for TM training. As we have demonstrated in Chapter 8, it is possible to exploit the parallel information given in the respective other language audio stream to bias the ASR

systems for improved transcription performance. In the proposed context, the resulting improved ASR performance directly affects the quality of the extracted training data. In the following, we speak of lightly supervised training or pSp supervised training whenever we apply biased ASR systems to create training data from pSp audio.

11.2 Data and Baseline Systems

We use the same En/Sp development and evaluation sets as in Chapter 10; dev05, dev06 and eval07. The pSp audio corpus is a subset of the pSp audio corpus from Chapter 10, consisting of 67 sessions from the time period 08Sep05-01Jun06. The supervised Spanish ST training data is limited to 10h of manually transcribed Spanish audio and to a parallel text corpus comprising 100k Spanish words, manually translated into English. Detailed data statistics for the training sets are shown in Table 11.1.

	transcriptions	parallel text	pSp
sent./utt. [k]	6.5	3.9	52.3
words [k]	79.6	100.0	751.8
audio [h]	10.0	N/A	91.7

Table 11.1: Data statistics: Spanish speech translation training data.

The MT decoder, MT training procedure and English ASR are identical to the ones described in Chapter 10. For training translation models from pSp audio of simultaneous interpretation, we rely on the more simple utterance alignment strategy of padding the target parallel speech snippets with 2 seconds. The baseline Spanish ASR system uses sub-phonetically tied three-state HMMs and features a single, speaker-independent decoding pass. The AM is trained on 10h Spanish EPPS data via three iterations of Viterbi training[1]. The 3-gram LM is estimated on 179.6k running words from the AM training data reference

[1]The training is bootstrapped with labels from the context-independent system SP_{c1} described in Chapter 10

transcriptions and the Spanish side of the parallel text corpus used for supervised TM training. In order to avoid high out-of-vocabulary rates, we use a large recognition dictionary with 74.2K pronunciation entries. This resource-limited Spanish ASR system yields WERs in the range of 26-27% on our data sets as shown in Table 11.2.

	English-to-Spanish			**Spanish**-to-English		
	dev05	dev06	eval07	dev05	dev06	eval07
base WER	13.1	13.9	12.2	26.1	26.9	27.1

Table 11.2: English and Spanish baseline system word error rates.

11.3 Parallel Speech Audio for ASR Model Training

Unsupervised acoustic model training relies on automatic transcriptions created with an initial ASR system. The success of unsupervised AM training usually depends strongly on the ability to exclude erroneous transcriptions from training. The common approach is to use word confidences for selecting transcriptions suitable for training. Lightly supervised AM training [37] refers to the case where some imperfect human transcriptions, for example closed-captions provided during television broadcasts, can be used to either bias the initial ASR system for improved transcription performance or to filter erroneous ASR hypotheses. We examine unsupervised AM training and lightly supervised AM training. We introduce light supervision with the help of pSp audio of simultaneous interpreters, as we introduced in Chapter 9.

To introduce light supervision based on English pSp audio for Spanish AM and LM training, we automatically translate the English parallel speech into Spanish and bias the Spanish ASR LM to prefer n-grams seen in the automatic translation. We distinguish between two different types of LM bias; a 'session bias' and an 'utterance bias'. Session bias refers to the case where we first automatically

translate the English audio of one complete European Parliament session into Spanish, and we then interpolate the baseline Spanish LM with a LM built on the automatic translation. Utterance bias, on the other hand, refers to the case where we bias the Spanish LM for each Spanish speech utterance. We achieve this by first translating the 6 seconds word padded English speech snippet. We then prefer the uni-grams found in the translated speech snippet, by boosting the baseline Spanish LM probability of these uni-grams, similar to a cache LM. The boosting of the uni-gram probability is realized by subtracting a discount value d from the (positive) LM log score of the current ASR hypothesis. The discount value d for a uni-gram u is estimated as follows:

$$d(u) = \begin{cases} w * LM_{score}(u) & \text{for } LM_{score}(u) \geq t \\ 0 & \text{for } LM_{score}(u) < t \end{cases}$$

with $LM_{score}(u)$ being the baseline LM score for the uni-gram u and weight w and threshold t estimated on dev05 via a grid search. Table 11.3 shows the influence of the session LM bias and the combination of utterance LM bias and session LM bias on the Spanish WER on dev05; the WER is reduced by 6% relative from 26.1% to 24.5%.

baseline	session bias	session & utterance bias
26.1	25.4	24.5

Table 11.3: Biasing ASR with parallel speech; Spanish word error rates on dev05.

For unsupervised and lightly supervised AM training, we utilize ASR word confidences in the following manner: speech frames associated with words that have an ASR word confidence of $c < 0.8$ are ignored; all other speech frames contribute to the training with a weight of 1. The value of c is estimated on our dev set. Training is realized via three iterations of Viterbi training. All iterations include 10h of manually transcribed audio plus 92h of automatically transcribed audio. Results obtained with the re-trained AMs are listed in Table 11.4, along

with results for unsupervised LM training. The first two columns of Table 11.4 specify if the baseline AM/LM was used or a model trained with the additional 92h of automatically transcribed Spanish speech. The case of light supervision during ASR decoding via a session+utterance bias is marked with a subscript $_b$. For example, the last row in the table refers to the case where we used the biased baseline ASR system to create additional AM and LM training data. The values shown in brackets represent the WER on dev05, when biasing the ASR with knowledge from the English parallel speech. Since we do not have English pSp available for dev06 and eval07, such a bias is not possible on these data sets. The results show that light supervision during training benefits ASR performance.

AM	LM	dev05	dev06	eval07
base	base	26.1 [24.5$_b$]	26.9	27.1
+92h	base	24.0 [23.0$_b$]	24.9	25.5
base	+92h	24.5 [23.3$_b$]	25.7	25.5
+92h	+92h	22.5 [21.5$_b$]	24.0	24.2
+92h$_b$	+92h$_b$	22.0 [21.6$_b$]	23.5	23.8

Table 11.4: Re-raining the Spanish acoustic model and language model with additional 92h of automatically transcribed parallel speech: influence on word error rate. Results marked with $_b$ were achieved by applying light supervision (session & utterance bias) during decoding.

In contrast to AM training, we do not utilize ASR word confidences during LM training. We estimate a LM on the Spanish ASR first-best hypotheses and interpolated this LM with the baseline LM. The interpolation weight is chosen to minimize the LM perplexity (PPL) on the dev set. Table 11.5 lists the PPL of the baseline LM and of the interpolated LMs, using transcriptions from the baseline and biased baseline Spanish ASR during training. The LM used to compute the dev05 PPLs (marked by *) does not include automatic transcriptions of dev05 itself. We found that, while the PPL decreases much more if ASR first best hypotheses of the same session are included in the LM, ASR transcription performance does not benefit due to an overly strong bias towards transcription

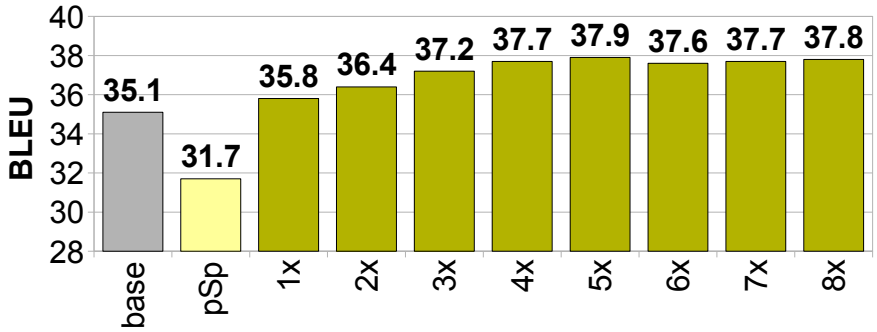

Figure 11.2: Combining parallel speech (pSp) training data with our baseline parallel text training corpus. The baseline training corpus of manual translation receives a higher weight by repeating it x times. Results are shown for Sp→En text translation on dev06.

errors made by the initial ASR. Therefore, whenever we automatically transcribe our pSp corpus with an ASR system that includes a re-trained LM, we use session-specific LMs that exclude ASR transcriptions of the very same session.

LM	dev05	dev06	eval07
base	182	269	276
+92h	129*	202	206
+92h$_b$	127*	200	204

Table 11.5: Language model (LM) re-training with additional 92h of automatically transcribed Spanish parallel speech: influence on perplexity.

11.4 Parallel Speech Audio for MT Model Training

When traditional MT training data is available, it is necessary to determine how traditionally trained translation models can benefit most from additional pSp audio. By simply extending our parallel text corpus of 100k manually translated words with the automatically transcribed pSp training data, and re-training the TM on this extended training corpus, we observe only small improvements on

dev06 of 0.7 BLEU points from 35.1 to 35.8. We therefore examine if a higher weighting of word alignments that stem from the supervised part of the combined corpus is helpful. The main idea is to aid the GIZA++ word alignment process on the pSp part. We achieve this higher weighting by simply duplicating the supervised training corpus x times. Figure 11.2 gives an overview of the Sp→En text translation (0% WER of the Spanish input) results on the dev set for $x \in \{1, 2, ..., 7, 8\}$. The figure also lists translation performance numbers in BLEU for the baseline TM, trained only on supervised training data, and for a TM trained only on the automatically transcribed and aligned pSp corpus. The best translation results are achieved by adding the supervised parallel text corpus of manual translations 5 times to the combined training corpus. However, it should be noted that the BLEU score variations observed for adding the pSp training data 3–8x times to the supervised training corpus are statistically not significant. Tables 11.6 and 11.7 list the achieved text translation results for both translation directions. The presented results are obtained with ASR transcriptions created with the baseline ASR systems.

TM	dev05	dev06	eval07
base	41.1	35.1	35.2
+92h	44.5	37.9	37.8

Table 11.6: Translation model (TM) re-training with additional 92h of automatically transcribed Spanish parallel speech: Sp→En text translation results in BLEU.

11.5 Speech Translation Results

In this section, we present our results for the complete ST chain of ASR and subsequent MT on the ASR first best hypotheses. We also pay special attention to the case of strong resource limitation, in which only 10h of transcribed Spanish AM data is available, but no baseline MT.

LM	TM	dev05	dev06	eval07
base	base	26.0	27.2	25.7
+92h	base	27.9	29.1	27.5
base	+92h	27.7	28.3	27.6
+92h	+92h	30.5	30.6	29.6

Table 11.7: Translation model (TM) and language model (LM) re-training with additional 92h of automatically transcribed Spanish parallel speech: En→Sp text translation results in BLEU.

Table 11.8 lists the speech translation results for En→Sp. We compare results of the baseline ST system with a ST system that includes unsupervised training data created with the baseline Spanish ASR. The eval set BLEU score increases by 3.2 points from 21.6 to 25.2 for the re-trained ST system. The case where no baseline automatic translation is possible due to the lack of parallel text data, is shown in the last row. With a TM trained solely on 92h of pSp audio, we achieve a translation performance of 19.9 BLEU points. In Table 11.9 we show speech translation results for Sp→En. Here, we examine two additional scenarios: first, we examine the effect of lightly supervised AM and LM training on the ST end result (row 3, entries marked with $_b$) and second, we address the effect of the improved transcription performance on TM training (last row). Specifically, the results in the last row of the table refer to the case where the pSp automatic transcriptions used for TM training came from the already re-trained ASR. All other listed results are achieved by using models that were re-trained with pSp transcriptions from either the baseline ASR or the biased baseline ASR. Re-training the ST models with baseline ASR transcriptions improves the eval BLEU score by 3.0 points from 25.3 to 28.3. Using ASR hypotheses from the biased Spanish ASR does not improve the overall speech translation result on our evaluation set, although ASR transcription performance is slightly improved, as shown in Section 11.4. In the scenario where no parallel text data for TM training is available, we achieve an eval BLEU score of 24.9—only slightly below the translation performance of the baseline system that is based on parallel text data. The translation performance of the pSp-only system can be further

increased by 0.7 BLEU points, when using the re-trained Spanish ASR system to transcribe the pSp corpus, instead of only using the baseline ASR system. This result suggests to introduce at least one iteration in the proposed training scheme, where ST models are first re-trained with transcriptions from the baseline ASR systems, and then, subsequently trained with transcriptions from systems that already benefit from re-trained models.

LM_{MT}	TM		dev05	dev06	eval07
base	base		24.0	22.4	21.6
+92h	+92h		28.5	25.7	25.2
+92h	92h		23.8	20.4	19.9

Table 11.8: Re-training with additional 92h of automatically transcribed Spanish parallel speech: En→Sp speech translation results in BLEU. The last row shows results achieved with a translation model purely trained from parallel speech (no baseline parallel text corpus).

AM	LM_{ASR}	TM		dev05	dev06	eval07
base	base	base		31.2	25.1	25.3
+92h	+92h	+92h		34.8	28.0	28.3
$+92h_b$	$+92h_b$	$+92h_b$		35.7	28.8	28.4
+92h	+92h	92h		31.8	24.2	24.9
+92h	+92h	$92h_{i=1}$		32.7	25.2	25.6

Table 11.9: Re-training with additional 92h of automatically transcribed Spanish parallel speech: Sp→En speech translation results in BLEU. Results marked with $_b$ were achieved by applying light supervision (session & utterance bias) during ASR decoding. The last two rows of the table list results achieved by only using parallel speech for translation model training (no baseline parallel text corpus). The results of the last row were achieved by applying the re-trained Spanish ASR system to the parallel speech audio for translation model training. All other results are based on parallel speech training data transcribed with the Spanish baseline ASR system.

11.6 Chapter Summary & Discussion

In previous chapters, we developed and examined approaches for (a) biasing ASR and MT with parallel speech audio and (b) training ASR and MT models with the help of parallel speech audio. With the framework introduced in this chapter for creating acoustic model, language model and translation model training data from parallel audio, we successfully tied the developed approaches together to significantly improve all major models involved in statistical speech translation. Specifically, we considered the scenario of ST between a resource-rich and a resource-limited language, and we reported significant performance improvements for the resource-limited ST models by enriching the limited training data resources with training data that was automatically created from parallel speecg audio.

11. INTERPRETATION AS SPEECH TRANSLATION TRAINING DATA

12

Speech Translation from Consecutive Interpretation

In the previous chapters we explored parallel speech audio as a novel data resource for training speech translation systems. We argue that the presented approaches are of special interest in the context of limited data resources. We consider the result that translation models can (a) be trained from scratch with parallel speech audio, and (b) be improved with parallel speech audio as additional training data, as one of our most important findings, as parallel text data is one of the ST training resources that is especially hard to acquire. However, our experiments remain limited to parallel speech audio of English/Spanish simultaneous interpretation. As we believe that consecutive interpretation is the prevailing from of interpretation in situations that ask for a rapid development of ST systems, we examine in this chapter if our findings regarding parallel speech trained translation models remain valid in the context of consecutive interpretation between English and the resource-limited language Pashto.

12.1 Previous Results & Chapter Outline

The experiments presented in Chapter 10 suggest that pSp-trained translation models mirror the training corpus size-dependent performance of parallel text trained translation models, just at a lower level. We estimated the 'yield' of English/Spanish SI audio to be around $n \cdot 10^{-1}$, meaning that we observed a pSp

corpus of n interpreted words to yield a similar translation performance in BLEU as a parallel text corpus of $n \cdot 10^{-1}$ translated words. In general, the yield of pSp audio certainly depends on different factors, as for example the type and 'quality' of used interpretation (CI vs. SI, as explained in Chapter 7), language pair and WERs of the ASR systems used to transcribe source and target language speech. However, assuming WER ranges as thus far considered, we hypothesize the general yield of parallel speech to be within the same order of magnitude as for English/Spanish SI, that is, somewhere in the range of $n \cdot 10^{-1}$, but certainly not $n \cdot 10^{-2}$. To further support this hypothesis, we examine the development of Pashto→English speech translation on the basis of pSp audio from consecutive interpretation. Specifically, we explore English/Pashto pSp audio as (a) the sole data source for TM training in Section 12.3, and (b) as additional training data, that is to be mixed with parallel text (Section 12.4).

12.2 Experimental Setup

12.2.1 US Darpa's TransTac project

Our experiments are based on data resources provided within US Darpa's TransTac project. The stated mission of TransTac is 'to demonstrate capabilities to rapidly develop and field two-way translation systems that enable speakers of different languages to spontaneously communicate with one another in real-world tactical situations'. One requirement of the program is to support new languages in less than 100 days. TransTac concentrates of languages of interest to national security. In different phases of the program, two-way ST was developed between (a) English and (b) languages like Iraqi, Farsi and Dari. The latest phase of the program concentrates on Pashto—a language spoken mostly in Afghanistan and western Pakistan. Typical scenarios considered within TransTac are in the form of interviews, where for example an English-speaking soldier interviews a Pashto-speaking Afghani.

During the course of the project, it became strongly apparent that the most pressing bottleneck in terms of rapid ST development and system performance is

the time-consuming (and costly) progress of transcribing and translating foreign language recordings. For a minimal domain mismatch, these recordings are collected from typical communication scenarios as they already occur in the field. Without ST solutions available, cross-lingual communication is typically achieved with the help of consecutive interpreters. An example for such a cross-lingual dialog is depicted in Figure 12.1. Each native speech utterance is accompanied by its CI utterance in the example. Further, a manual translation of the non-English parallel speech is provided.

Interviewer:	what is it that you wanted to speak with me about today
Interpreter:	تشکر زه ښه یم *thanks I am fine*
	تاسو نن زما سر ه د څه شي په باره کښې خبرې کولې w*hat do you want to talk about with me today*
Respondent:	ما غوښتل تاسو سر ه وغږېږم دلته بعضې شیان دې دلته د تېلو ځای دي د *I wanted to talk to you --* *there are some things here in the oil station that I want to talk to you about*
Interpreter:	I just want to talk with you about -- there is a -- a gas station
	I would like to talk about that with you
Interviewer:	okay and what is the importance of this gas station
Interpreter:	ښځي صحیح ده د دې په هکله تا څه غوښتل چې زما سر ه ووایې *it is okay - what do you want to tell me about this*

Figure 12.1: Consecutive interpretation example.

12.2.2 Data and Scoring

Only very limited amounts of data resources are available for English/Pashto ST development. Table 12.1 lists the statistics of our English/Pashto parallel speech corpus. It shows the amount of native speech (English interviewer, Pashto respondent) and interpreter speech in hours of audio and number of uttered words. In contrast to our English/Spanish pSp corpus, we have manual reference transcriptions available. Further, we have manual reference translations for each native speech utterance available. In addition to the pSp corpus, we make use of a 'traditional' English/Pashto parallel text corpus of manual translations. This parallel text corpus has 12.4k translated Pashto respondent utterances. The Pashto part

	Native		Interpr.	
	En	Pa	En	Pa
audio [h]	23.0	25.2	26.7	29.5
words [k]	358	374	333	399

Table 12.1: English/Pashto parallel speech audio statistics.

	Pa→En	
	dev	eval
audio [min]	45.8	24.0
words [k]	6.7	3.6

Table 12.2: Pashto→English development and test set.

comprises 260k words; the English part has 214k words. Table 12.2 lists the statistics of the Pashto→English development and test sets. Both sets are based on native Pashto respondent speech and feature only one reference translation for scoring (IBM BLEU).

12.2.3 Sentence Segmentation

In order to utilize the English/Pashto parallel speech audio corpus in our standard TM training setup, we have to create a sentence-aligned bilingual text corpus first. We can exploit the fact that each speaker takes turns in consecutive interpretation, with each speaker producing only a few utterances in each turn. To introduce speaker-turn-based sentence alignment, we rely on the manual utterance segmentation and the role description (interviewer, respondent, interpreter) found the transcription files. As the interviewer speech is recorded on a different audio channel from the interviewer/respondent speech, we argue that an algorithm that is based on automatic utterance segmentation and automatic speaker identification will provide a very similar performance. All of our training runs are based on aligned speaker turns, even when manual translations are used for model building. This is possible, since each speech utterance is accompanied with

a manual translation in the corpus. Our decoding/scoring runs on the development and evaluation sets observe the speech utterance segmentation.

12.3 Consecutive Interpretation as Only Data Source

In a situation where only untranscribed parallel speech audio is available, the minimal requirement for speech translation development are two ASR systems to enable the automatic transcription of source and target language speech. In the case of ST development between a resource-rich and a resource-deficient language, ASR systems for the resource-rich language may already be available. In our case, we have an in-domain English ASR system from previous phases of the TransTac project available[1], as previous phases considered ST between (a) English and (b) Iraqi, Farsi and Dari. However, we have no pre-existing Pashto ASR at hand. To enable Pashto→English speech translation and to be able to automatically transcribe additional parallel speech audio, we train a Pashto ASR system on the 25.2h of Pashto respondent speech found in our pSp corpus. For AM and LM training, we rely on the manual transcription of this respondent speech (374k words). Table 12.3 lists the English and Pashto WER and LM perplexity for the automatically transcribed parts of the pSp corpus. The interpreter speech frequently suffers from a heavy foreign accent, explaining the significantly higher WER on interpreter speech compared to native speech. The Pashto WER on the Pashto→English development and test set is 33.7% and 33.9%, respectively. The LM perplexity is 157 and 148, respectively.

To examine our hypothesis that n interpreted Pashto words yield approximately the same translation performance, measured in BLEU, as $n \cdot 10^{-1}$ translated words, we examine three different systems estimated on the parallel speech corpus. System A uses translation models trained on the manually transcribed

[1]English and Pashto ASR both feature only one decoding pass and small models tuned to the real-time requirements of TransTac evaluations. Both systems were developed by the author using a training setup similar to the one described in Chapter 5. The English ASR includes discriminative AM training in the form of boosted MMI training [59].

	Native		Interpr.	
	En	Pa	En	Pa
PPL	68	-	75	196
WER	16.3	-	30.7	44.9

Table 12.3: Parallel speech audio: language model perplexity (PPL) and word error rate.

and translated Pashto respondent speech that is present in the parallel speech corpus. In System B the English translations are replaced by the manual transcription of the interpreter speech. System C finally uses the English automatic transcription (30.7% WER) of English interpreter speech. While system C does not suffer from word errors on the Pashto side (we use the Pashto reference transcription), the English word error rate is on the same level as the worst WER level we considered English/Spanish parallel speech audio. Table 12.4 lists the text and speech translation performance in BLEU for all three systems. Table 12.5 lists the English type and token coverage of the training corpora A and B in regard to dev, showing that corpus coverage does not play an important role. As we expect system B and C to perform on the same level as a system that is trained on approximately 40k manually translated words, we compute the corpus-size dependent text translation performance of system A for increments of 10k words, until system A meets the performance of system B. The result is depicted in Figure 12.2. It shows that our prediction was accurate. Another important hypothesis is that the translation performance of parallel speech trained translation models mirrors the translation performance of parallel text trained translation models, just at a lower level. While only a very limited amount of English/Pashto parallel speech audio is available, we still can observe the same trend as observed for English/Spanish parallel speech audio, when we compute the corpus-size dependent text translation performance of system A and B in increments of 90k words, compare Figure 12.3.

	text		speech	
	dev	eval	dev	eval
A	17.6	17.8	14.6	15.2
B	11.8	13.0	10.5	10.0
C	10.9	10.5	9.4	10.2

Table 12.4: Pashto→English text and speech translation performance for systems A, B and C. The Pashto part of the parallel translation model training corpus consists of manually transcribed Pashto respondent speech. The English part consist of (A) manual English translations; (B) manually transcribed interpreter speech or (C) automatically transcribed (30.7% word error rate) interpreter speech.

	token	type
A	98.8	92.9
B	98.1	90.0

Table 12.5: Vocabulary and corpus coverage for systems A and B.

12.4 Consecutive Interpretation as Additional Source

To further examine the value of pSp audio as TM training data in addition to parallel text, we estimate a TM on the parallel text corpus of 260k translated Pashto words. We refer to the system using this TM as system D. We then increase the parallel text corpus with the training corpus of system A, B or C and estimate new translation models, resulting in systems D+A, D+B and D+C. Table 12.6 gives an overview of the text and speech translation performance of these systems. With English and Pashto ASR available, it is possible to automatically transcribe more pSp audio, promising further gains in translation performance at a relative low cost. For example, we can automatically transcribe the part of the pSp corpus formed by English interviewer speech (16.3% WER) and respective Pashto interpretation (44.9% WER)—referred to in the following as training data F. Despite the very high Pashto WER, we achieve further gains in text and speech translation performance by adding training data F to D+C, as shown in the last

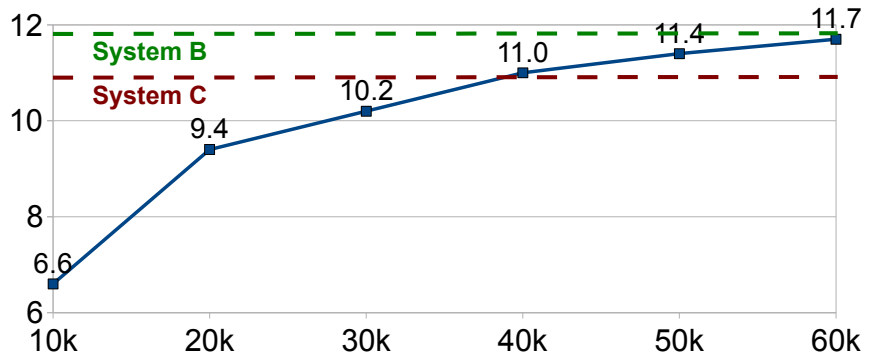

Figure 12.2: Corpus-size dependent BLEU score on dev of system A (trained on manual translation).

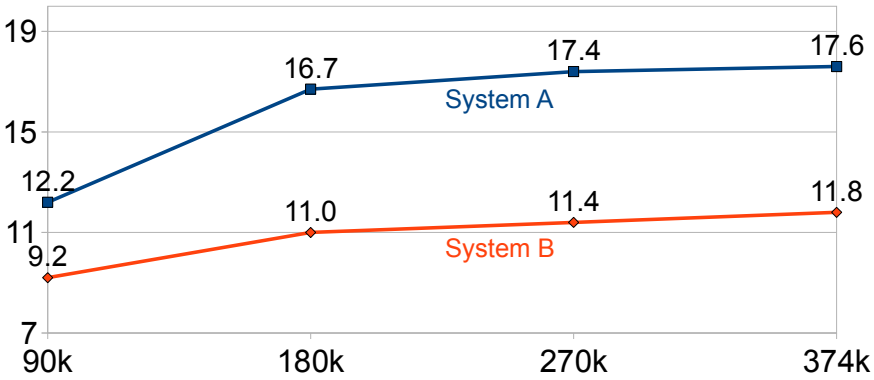

Figure 12.3: Corpus-size dependent BLEU scores on dev of system A (trained on manual translation) and B (trained on manually transcribed interpretation).

row of Table 12.6. The observed improvements for system D+C+F compared to system D are statistically significant ($p < 0.05$). Similar to the experiment described in Section 11.4, these results are achieved by weighting training data D+C and training data F differently. In the case of text (speech) translation, D+C was repeated 3 (4) times in the final training corpus D+C+F.

12.5 Chapter Summary & Discussion

In this chapter we have shown that our findings regarding parallel speech trained translation models, made in the context of English/Spanish simultaneous in-

	text		speech	
	dev	eval	dev	eval
D	**12.3**	**12.3**	**11.2**	**10.0**
D+A	18.4	17.5	16.0	14.2
D+B	14.6	14.7	12.7	12.2
D+C	13.8	13.4	11.6	12.0
D+C+F	**14.7**	**14.9**	**12.5**	**12.4**

Table 12.6: Increasing translation performance by adding more training data. Baseline parallel text training corpus (D) plus (A) more manual translations; (B) manually transcribed parallel speech audio or (C) automatically transcribed parallel speech audio. Training corpora A, B and C consist of either translated or interpreted Pashto respondent speech. Training corpus F consists of automatically transcribed parallel speech formed by interpreted English interviewer speech.

terpretation, remain valid in the context of consecutive interpretation between English and the resource-limited language Pashto. We observed a similar yield of interpretation audio compared to parallel text in terms of BLEU score for English/Pashto consecutive interpretation. Further, we reported statistically significant improvements in BLEU metric by enhancing parallel text with the parallel speech audio of consecutive interpretation for translation model training. These results further support our hypothesis that automatically transcribed parallel speech audio can present a valuable, low-cost data resource for speech translation development.

12. SPEECH TRANSLATION FROM CONSECUTIVE INTERPRETATION

13

Results and Discussion

Despite the continuously increasing demand for automatic solutions that support cross-lingual, verbal communication between myriad languages, the development of deployable speech translation systems continues to be viable for a mere handful of languages. A combination of unsolved research challenges in speech translation (ST), including insufficient quality of output and high development cost, are responsible for this undesirable situation. This thesis has described several contributions aimed to resolve this situation.

First, we introduced a sentence segmentation and punctuation recovery scheme for speech translation. This scheme helps to improve automatic translation of spoken language by targeting the mismatch between automatic speech recognition (ASR) system output and 'traditional' machine translation (MT) training data (presented in the form of sentence-aligned bilingual text of manual translations). By applying this sentence segmentation and punctuation recovery scheme, we showed significant improvements in translation performance, measured in BLEU, for three very different spoken language translation tasks: English→Spanish translation of speeches given in the European Parliament as well as Chinese→English and Arabic→English translation of broadcast news.

Further, we introduced an approach that supports the cost-effective development of automatic speech recognition systems in the various languages of the European Union. This approach exploits the freely available data resources given in

the context of European Parliament Plenary Session: the live broadcast speeches along with their simultaneous interpretations, and the Parliamentary proceedings, published on the Parliament's web pages. By exploiting these data resources, the presented approach enables the training of acoustic models without having costly verbatim transcriptions available.

Finally, we introduced a new, cost-effective way of training speech translation systems from parallel speech (pSp) audio: audio recordings of interpreter-mediated communication. We developed various approaches that enable: (a) the automatic extraction of ST training data from such audio recordings; and (b) successful exploitation of this ST training data. Thus, we are able to significantly reduce the amount of costly human supervision that has typically characterized speech translation system development. Specifically, we have shown that all major statistical models involved in state-of-the-art speech translation (acoustic models, language models and translation models) can benefit from parallel speech audio by applying unsupervised and pSp-supervised training techniques. In particular, we have also shown that translation models can be trained from scratch, without any parallel text data of sentence-aligned manual translations, by using automatically transcribed parallel speech. In our experiments, we covered both forms of (speech) interpretation: simultaneous interpretation (SI) and consecutive interpretation (CI). While we developed our approaches in the context of a resource-rich task, we payed special attention to the situation of data resource limitation, as we argue that training ST from interpretation is of special value in the context of resource limitation. Our results have shown that the approach is robust against low automatic transcription performance, confirming the approach's feasibility in the context of resource limitation. We consider the result that parallel speech audio can replace as well as enhance parallel text data as a training resource for translation model development as one of our most important findings, as domain specific parallel text is especially hard to come by and costly to create.

One problem that was omitted by us is the fact that additional parallel speech potentially includes high amounts of out-of-vocabulary words. The magnitude of

the out-of-vocabulary problem depends strongly on the domain and the ASR system vocabulary size, as for example shown by Hetherington [27]. Optimizing the vocabulary size to the current domain is one solution, but this may be hard to accomplish in the context of ASR in resource-limited languages. Approaches that automatically identify unknown words and use phoneme or grapheme representation of these words are another possibility. A somewhat related problem is the question of how to address ST development in the context of languages that do not have an acknowledged written form. Besacier et al. [6] propose an interesting solution to this problem. They propose to apply phone-based ST where translation models are learned on a parallel corpus of foreign phone sequences and corresponding English translation. It is also possible to utilize such a parallel training corpus of English word sequences and foreign phoneme sequences for the task of automatically discovering new word units for ASR, as shown by Stüker et al. [73].

We believe that the specific mixing strategy for optimally combining training data with parallel speech data, at an acceptable cost level deserves more attention. For example, one could ask if it makes more sense to manually transcribe higher amounts of speech data for reduced word error rate on the resource-deficient language or if it is more helpful to manually translate data. In this context, approaches that automatically identify interpretation ASR hypotheses that are problematic in terms of word error rate or content are of special interest.

We further believe that future work has to address larger amounts of parallel speech audio and more language pairs, to further support our hypotheses regarding the translation performance of pSp-trained translation models. While the attached collection effort of additional parallel speech audio may be considered the biggest obstacle, one has to realize that: (a) interpretation happens daily on a massive scale; (b) simultaneous interpretation typically involves considerable amounts of equipment (sound proof booths, etc.) that directly enable the recording of parallel speech audio; and (c) that huge amounts of money flow into the development of ST systems for CI-like situations. The latter point implies that there are many consecutive interpretation situations in which the recording of source and target language speech is feasible. Therefore, our results promise

substantial improvements in automatic translation of text and speech, achieved at a relatively low additional cost, by collecting more parallel speech audio.

Appendix A

Kurzfassung in Deutscher Sprache

A.1 Automatische Sprachübersetzung

Automatische Sprachübersetzung basiert auf der Kombination zweier Technologien: automatische Spracherkennung (automatic speech recognition, ASR) und maschinelle Übersetzung (machine translation, MT) von geschriebenem Text. Sprachübersetzung kann hierbei als das Problem betrachtet werden, ASR und MT in einer mit der Leistung heutiger Computer machbaren Art und Weise zu kombinieren, so dass eine bestmögliche Übersetzungsleistung auf der fehlerbehafteten Ausgabe automatischer Spracherkennungssysteme erzielt wird. Ein Problem bei der Intergration von ASR und MT ist die Diskrepanz zwischen der Ausgabe automaticher Spracherkennung und dem Format der Trainingsdaten die überlicherweise zum Trainieren maschineller Übersetzungssysteme zur Verfügung stehen. Die Ausgabe von ASR Systemen beinhaltet Erkennungsfehler, ist nicht mit Satzzeichen versehen und unerliegt überlicherweise einer Segmentierung die auf Regionen im Sprachsignal beruht, in denen keine Sprache detektiert werden kann. Letzteres hat insbesondere zur Konsequenz, dass die Segmentierung der Spracherkennerausgabe einer Segmentierung auf Satzgrenzen hin nicht ähnlich ist. Trainingsdaten maschineller Übersetzungssysteme hingegen bestehen typischerweise aus manuell übersetzten Texten geschriebener Sprache (im Gegensatz zu gesprochener Sprache) mit korrekten Satzsegmenten und Satzzeichen.

Im Rahmen dieser Dissertation wird ein Verfahren zur automatischen Satzsegmentierung und Punktuation für Sprachübersetzungsysteme eingeführt, welches diese Diskrepanz in Angriff nimmt. Indem wir (a) modifizierte Übersetzungsmodelle zum impliziten Einfügen von Kommata während der Übersetzung einführen und (b) ein auf einem Entscheidungsbaum basierendes Verfahren zur Satzsegmentierung von Spracherkennungsausgabe entwickeln, erzielen wir signifikante Verbesserungen in der Übersetzungleistung von Sprachübersetzungssystemen für drei sehr unterschiedliche Sprachenpaare.

A.1.1 Statistische Modelle in der Sprachübersetzung

Automatische Spracherkennung und maschinelle Übersetzung werden heutzutage von statistischen Modellierungsansätzen dominiert. Während diese statistischen Modellierungsansätze hauptsächlich zu den beachtlichen Leistungssteigerungen automatischer Sprachübersetzungsysteme in den letzten Jahrzehnten beigetragen haben, sind die enormen Ansprüche an Trainingsdaten (und die damit verbundenen Kosten) dieser Verfahren einer der Hauptgründe weshalb die Entwicklung von einsatzfähigen Sprachübersetzungssystemen auf nur eine handvoll von Sprachenpaaren begrenzt bleibt. Betrachtet man die statistischen Modelle die an automatischer Sprachübersetzung beteiligt sind—akustisches Modell und Sprachmodell der Quellsprache für die ASR sowie Übesetzungmodell und Sprachmodell der Zielsprache für die MT—so lässt sich das Übesetzungsmodell als das kostenintensivste Modell identifizieren. Während domänenspezifische monolinguale Ressourcen—transkribierte Sprachaufnahmen zum Trainieren akustischer Modelle sowie Textkorpora zum Trainieren von Sprachmodellen—schwer beschaffbar sein können, sind domänenspezifische bilinguale Textdaten, bestehend aus satzalignierten manuellen Übersetzungen, noch seltener und kostenintensiver zu erstellen. Diese Arbeit hat zum Ziel die kostspielige menschliche Überwachung, welche sich durch die Trainingsdatenanforderungen modernen Sprachübersetzungssysteme ergeben, mit Hilfe einer oftmals zur Verfügung stehenden, jedoch bislang ignorierten Ressource zu limitieren. Im Detail wird angestrebt, automatische

Sprach**übersetzung** mit Hilfe von **Audioaufnahmen** menschlicher **Interpretation**sszenarien zu trainieren.

A.2 Moderne Sprachübersetzungsysteme und Menschliche Interpretation

Mit dem Ziel Anwendungsszenarien zu identifizieren und Methoden zu entwickeln die es erlauben menschliche Interpretation für automatische Sprachübersetzung auszunutzen, betrachten wir zunächst die Entwicklung eines modernen Sprachübersetzungsystems und dessen Kerntechnologien im Kontext von Sitzungen des europäischen Parlamentes. Im Detail betrachten wir die Entwicklung von automatischer Spracherkennung und maschineller Übersetzung (und deren Kombination) für den enormen Transkriptions- und Übersetzungsaufwand der für die Erstellung der Sitzungsprotokolle ("final text editions"), in den verschiedenen Sprachen der Europäischen Union notwendig ist; vergleiche auch Abbildung A.1.

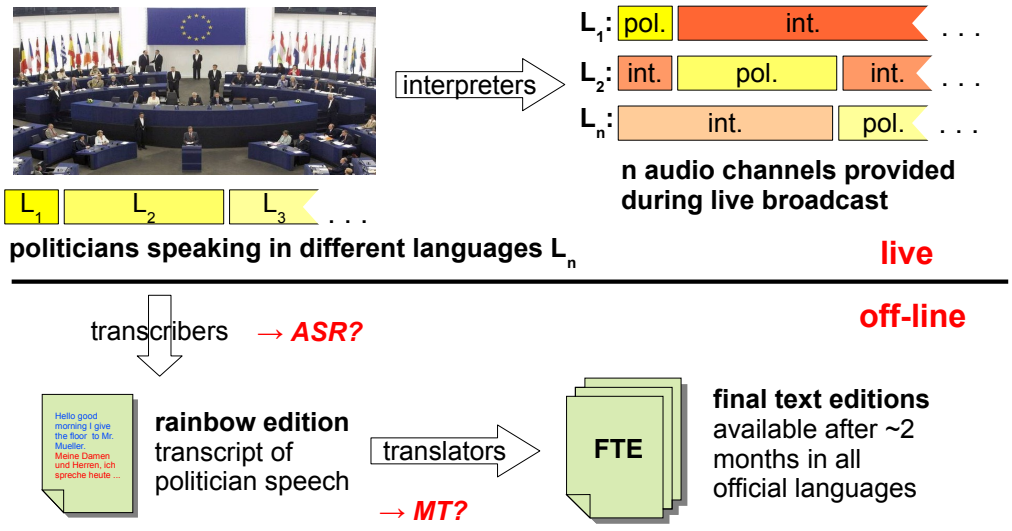

Figure A.1: Sitzungen des Europaparlaments und der damit verbundene Transkriptions- und Übersetzungsaufwand.

Indem wir die während der Sitzungen im Europaparlament angebotenen men-

schlichen Simultanübersetzungen (Interpretationen) ausnutzen um die den ASR und MT Systemen unterliegenden statistischen Modelle zu adaptieren, erzielen wir eine gesteigerte Erkennungs- und Übersetzungleistung dieser Systeme. Die Übersetzungsleistung der MT kann hierbei direkt durch automatische Transkriptionen von Interpretationen in der jeweiligen Zielsprache gesteigert werden. Die Adaption der ASR Modelle in der jeweiligen Quellsprache wird mit Hilfe von automatischen Sprachübersetzungen von Interpretation(en) in einer (oder mehreren) Zielsprache(n) zurück in die Quellsprache bewerkstelligt. Die Leistungssteigerungen werden trotz der fundamentaler Unterschiede (Abbildung A.2) die zwischen Interpretation ("paraller Sprache") und Übersetzung herrschen, erzielt.

Simultaneous Interpretation	Consecutive Interpretation
SPANISH UTTERANCE: "trataremos de que todo el personal tenga" *TRANSLATION: "we shall try that all the staff will get"* PARALLEL SPEECH: "... in addition to that we are going to try to make sure that members of staff from different members states of the european union will be granted an equal status ..."	ENGLISH UTTERANCE: okay and what is the importance of this gas station PARALLEL SPEECH: بيخي صحىح ده د دي په هکله ات خه غوښتل چي زما سر هوواىي *TRANSLATION: it is okay - what do you want to tell me about this*

Figure A.2: Unterschiede zwischen Übersetzung (translation) und Interpretation ("parallel speech").

A.3 Trainieren von Sprachübersetzungsystemen aus Audioaufnahmen Menschlicher Interpretation

Das in Abschnitt A.2 beschriebene Anwendungsszenario beruht auf (a) signifikanten Mengen von überwachten Trainingsdaten um ASR und MT Systeme trainieren zu können und (b) der Tatsache, dass Sprachaufnahmen von menschlichen Interpretern parallel zu exakt denselben Sprachaufnahmen der Quellsprache zur Verfügung stehen, für die Transkriptionen und Übersetzungen erzeugt werden

sollen. Jedoch sind viele der realen Anwendungsszenarien für Sprachübersetzung von einem Mangel an überwachten Trainingsdaten gekennzeichnet. Auch ist es für die meisten Anwendungsszenarien unwahrscheinlich anzunehmen, dass Sprachübersetzung angewandt werden soll wenn bereits menschliche Interpretation zur Verfügung steht. Aus diesem Grund entwickeln wir Methoden um Sprachaufnahmen von "paralleler Sprache" (Audioaufnahmen des Sprecher in der Quellsprache zusammen mit Audioaufnahmen des menschlichen Interpreters in der Zielsprache) für das Trainieren der an der automatischen Sprachübersetzung beteiligten statistischen Modelle auszunutzen. Solche trainierten Modelle ermöglichen letztendlich automatische Sprachübersetzung in Situationen in denen keine menschlicher Interpreter zur Verfügung stehen.

A.3.1 Trainieren von Akustischen Modellen im Kontext von Sitzungen des Europaparlaments

Obwohl enorme Mengen an Datenressourcen im Zusammenhang von Sitzungen des Europaparlaments zur Verfügung stehen zeigt sich auch schon hier ein gewisser Mangel an überwachten Trainingsdaten. Zwar stehen Unmengen an manuell übersetzten Textdaten in der Form von Sitzungsprotokollen der vergangenen Jahre in den verschiedenen Sprachen der Europäischen Union zur Verfügung, jedoch gibt es nur begrenzte Mengen an wortgetreuen Transkriptionen die sich für das Trainieren von akustischen Modellen eignen. Die manuell erstellten Sitzungsprotokolle weichen zum Teil erheblich von wortgetreuen Transkriptionen der im Parlament gehaltenen Reden ab. Zum einen werden die Sitzungsprotokolle mit dem Ziel einer bestmöglichen Lesbarkeit erstellt, zum anderen ist es den Rednern erlaubt, die Sitzungsprotokolle im Nachhinein abzuändern. Des weiteren werden keinerlei manuellen Transkriptionen/Protokolle der Simultanübersetzungen im Parlament erstellt. Inbesondere bleibt hierdurch die Entwicklung von Sprachübersetzungsystemen auf das Sprachenpaar Englisch/Spanisch begrenzt, da es nur hierfür signifikante Mengen an wortgetreuen Transkription gibt[1].

[1]Englische und spanische Transkriptionen wurden im Zusammenhang mit dem europäischen Projeckt TC-STAR erstellt

Um die Entwicklung von automatischen Sprachübersetzungsystemen in all den Sprachen der Europäischen Union zu unterstützen werden im Rahmen dieser Dissertation Verfahren zum Trainieren von akustischen Modellen untersucht, die lediglich mit einer "leichten" menschlichen Überwachung auskomment ('light supervision'). Das Training der akustichen Modelle basiert hierbei auf Audioaufnahmen von den für Europarlamentssitzungen live per Satellit übertragenen Audiospuren (siehe Abbildung A.1, oben rechts), sowie auf Informationen die aus bereits vorhandenen Sitzungsprotokollen automatisch extrahiert werden können. Leicht überwachtes Training von akustischen Modellen für die Sprache L_i und dazugehörigen Audioaufnahmen von Rednern im Parlament sowie von den Simultanübersetzern wird durch eine Adaption des ASR_i Sprachmodells mit Informationen erreicht, welche automatisch extrahiert werden aus (a) den Sitzungsprotokollen in der Sprache L_i, und (b) aus der parallelen Sprache die sich in den Audiokanälen für die Sprachen $L_{j\neq i}$ finden lässt.

A.3.2 Trainieren von Übersetzungsmodellen mit Hilfe von Interpretationen

Um parallele Sprache für das Trainieren von phrasen-basierten Übersetzungsmodellen verwenden zu können ist es zunächst notwendig die Sprache des Sprechers in der Quellsprache sowie die Sprache des Interpreters in der Zielsprache zu transkribieren. Wir verwenden hierfür automatische Spracherkennungssysteme. Die Hypothesen der Spracherkenner müssen dann aligniert werden, damit diese in einem standard Trainingssetup für phrasen-basierte Übersetzungsmodelle verwendet werden können. Hierfür entwickeln wir im Rahmen dieser Dissertation spezielle Verfahren die auf paralleler Sprache von Simultanübersetzungen oder konsekutiven Interpretationen zugeschnitten sind.

Da das Trainieren von Übersetzungsmodellen mit Hilfe von Interpretationen speziell für die Entwicklung von Sprachübersetzungsystemen im Kontext von Sprachen mit nur wenigen Ressourcen von Interesse ist, untersuchen wir dieses Verfahren zunächst auf den für die Sitzungen im Europaparlament zur

Verfügung stehenden Englisch/spanischen Simultanübersetzungen. Dies erlaubt es uns künstlich verschiedene Grade einer Ressourcenlimitation einzuführen und deren Effekt auf Übersetzungsmodelle die aus paraller Sprache trainiert wurden zu untersuchen. Die auf diese Art und Weise entwickelten Methoden zum Trainieren von Übersetzungsmodellen aus paraller Sprache werden erfolgreich im Zusammenhang einer echten Ressourcenlimitation angewandt. Konkret werden Übersetzungsmodellen aus der parallen Sprache von konsekutiven Interpretationsszenarien für das Sprachenpaar English/Pashto trainiert.

Unsere Experimente zeigen dass sich überraschend gute automatische Übersetzungresultate mit aus paraller Sprache trainierten Übersetzungsmodellen erreichen lassen. Im Zusammenhang von Englisch und spanischen Simultanübersetzungen zeigen wir zum Beispiel dass lediglich ein um den Faktor 10 grösserer Trainingskorpus aus **automatisch transkribierter** paraller Sprache notwendig ist, um dieselbe Übersetzungsleistung wie mit Modellen zu erzielen, die aus parallen, manuell übersetzten Texten trainiert wurden. Des weiteren zeigen unsere Experimente, dass sich parallele Sprache erfolgreich mit traditionellen Trainingskorpora, d.h. manuell übersetztenTexten, kombinieren lässt um eine gesteigerte Übersetzungsleistung zu erzielen.

A.3.3 Parallele Sprache als Trainingsressource für automatische Sprachübersetzung

Die entwickelten Methoden zum (a) adaptieren von ASR (und MT) Systemen mit Hilfe von paraller Sprache und (b) trainieren von akustischen Modellen und Übersetzungsmodellen aus paralleler Sprache lassen sich nun kombinieren um den enormen Aufwand an kostspieliger menschlicher Überwachung, welcher bisher für das Trainieren von Sprachübersetzungsystemen notwendig war, zu reduzieren. Abbildung A.3 zeigt unser System für das vollautomatische Extrahieren von Trainingsdaten aus paralleler Sprache. Wir untersuchen dieses Setup im Zusammenhang von Sprachübersetzung zwischen Sprachenpaaren in denen einer der beiden Sprachen durch einen Ressourcenmangel gekennzeichnet ist. Ziel ist es zusätzliche Trainingsdaten aus paralleler Sprache zu extrahieren, so dass die

statistischen Modelle, welche unter dem Ressourcenmangel leiden (in Abbildung A.3 farblich gekennzeichnet), verbessert werden können.

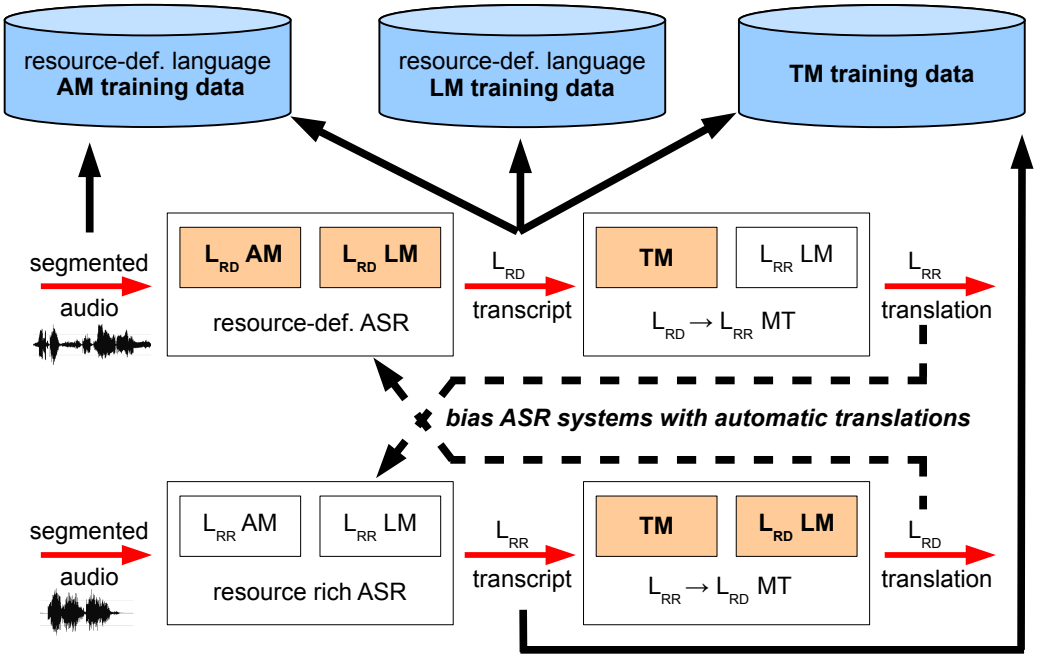

Figure A.3: Extrahieren von Trainingsdaten aus paralleler Sprache.

Die Kernkomponenten in dem Setup sind die beiden Spracherkennungssysteme. Es werden nur sehr kleine, auf wenigen Daten trainierte, anfängliche Spracherkennungssysteme benötigt. Insbesonderen kann ebenfalls mit nur kleinen anfänglichen maschinellen Übersetzungssystemen gearbeitet werden oder aber auch ganz auf anfänglichen Übersetzungssysteme verzichtet werden. Die Hypothesen der Spracherkennungssysteme, zusammen mit der Audioeingabe, lassen sich zum unüberwachten Trainieren von akustischen Modellen nutzen. Die Hypothesen sind desweiteren nützlich für das unüberwachte Trainieren von Sprachmodellen. Mit Hilfe der Methoden zum Trainieren von Übersetzungsmodellen aus paralleler Sprache lassen sich aus den Spracherkennerhypothesen letztendlich auch Übersetzungsmodelle erstellen. Die parallele Information die in dem Sprachsignal der jeweils anderen Sprache vorzufinden ist wird ausgenutzt, um die beiden Spracherkennungssysteme für eine gesteigerte Erkennungsleistung zu adaptieren.

Die gesteigerte Erkennungsleistung beeinflusst damit direkt die Qualität der auf diese Art und Weise automatisch erstellten Trainingsdaten.

References

[1] R. Al-Kahnjii, S. El-shiyab, and R. Hussein. **On the Use of Compensatory Strategies in Simultaneous Interpretation**. *Meta: Journal des traducteurs*, **45(3)**:544–557, 2000. 55

[2] Y. Al-Onaizan and L. Mangu. **Arabic ASR and MT Integration For GALE**. In *ICASSP*, Hawaii, USA, April 2007. 40, 43, 44

[3] A.Reddy and R. Rose. **Towards Domain Independence in Machine Aided Human Translation**. In *Interspeech*, Brisbane, Australia, September 2008. 5

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. **A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains**. *The Annals of Mathematical Statistics*, **41 (1)**:164–171, 1970. 14

[5] N. Bertoldi and M. Frederico. **A New Decoder for Spoken Language Translation Based on Confusion Networks**. In *ASRU*, San Juan, Puerto Rico, December 2005. 19, 65

[6] L. Besacier, B. Zhou, and Y.Gao. **Towards Speech Translation of Non Written Languages**. In *SLT*, Palm Beach, Aruba, March 2006. 125

[7] F. Van Besien. **Anticipation in Simultaneous Interpretation**. *Meta: Journal des traducteurs*, **44(2)**:250–259, 1999. 54

[8] J. Brousseaux, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon. **French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project**. In *Eurospeech*, Madrid, Spain, September 1995. 5

[9] P. Brown, S. Chen, S. DellaPietra, V. DellaPietra, A. Kehler, and R. Mercer. **Automatic Speech Recognition in Machine Aided Translation**. *Computer Speech and Language*, **8**(3):177–87, 1994. 5, 58

[10] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, , and R. L. Mercer. **The Mathematics of Statistical Machine Translation: Parameter Estimation**. *Computational Linguistics*, **19**:263–311, June 1993. 15

[11] F. Casacuberta, M. Federico, H. Ney, and E. Vidal. **Recent Efforts in Spoken Language Translation**. *IEEE Signal Processing Magazine*, **25 (3)**:80–88, May 2008. 15, 17, 18, 19

[12] C. J. Chen. **Speech Recognition with Automatic Punctuation**. In *Eurospeech*, Budapest, Hungary, September 1999. 40

[13] A.P. Dempster. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society*, **39 (1)**:1–38, 1970. 14

REFERENCES

[14] M. Dymetman, J. Brousseaux, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon. **Towards an Automatic Dictation System for Translators: the Transtalk Project**. In *ICSLP*, Yokohama, Japan, September 1994. 58

[15] M. Eck. **Developing Deployable Spoken Language Translation Systems given Limited Resources**. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2008. 8

[16] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. **The Karlsruhe-Verbmobil Speech Recognition Engine**. In *ICASSP*, Munich, Germany, April 1997. 14

[17] J.G. Fiscus. **A post-processing system to yield reduced word error rates:Recognizer Output Voting Error Reduction (ROVER)**. In *ASRU*, Santa Barbara, CA, USA, December 1997. 25, 36

[18] G. Foster, R. Kuhn, and H. Johnson. **Phrasetable Smoothing for Statistical Machine Translation**. In *Empirical Methods in Natural Language Processing*, Sydney, Australia, July 22-23 2006. 32, 33

[19] P. Fung. **A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora**. In *Parallel Text Processing*, pages 1–17. Springer, 1998. 7

[20] P. Fung and T. Schultz. **Multilingual Spoken Language Processing – Challenges for Multilingual Systems**. *IEEE Signal Processing Magazine*, **25 (3)**:89–97, May 2008. 19

[21] GALE. **Global Autonomous Language Exploitation**. http://www.darpa.mil/ipto/programs/gale/. 41

[22] M.J.F. Gales. **Maximum Likelihood Linear Transformation for HMM-based Speech Recognition**. *Computer Speech and Language*, **12**:75–98, 1997. 30

[23] M.J.F. Gales. **Semi-tied Covariance Matrices for Hidden Markov Models**. Technical report, Engineering Department, Cambridge University, Cambridge, England, February 1998. 30

[24] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney. **Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus**. In *ICASSP*, pages 825–828, Philadelphia, USA, March 2005. 21, 24, 25, 35, 76, 77

[25] Y. Gotoh and S. Renals. **Sentence Boundary Detection in Broadcast Speech Transcripts**. In *ISCA Workshop: ASR2000*, Paris, France, September 2000. 40

[26] Y. Gotoh and S. Renals. **Optimizing Sentence Segmentation for Spoken Language Translation**. In *Interspeech*, Antwerp, Belgium, August 2007. 40

[27] I. L. Hetherington. **A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding**. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1995. 125

[28] J. Huang and G. Zweig. **Maximum Entropy Model for Punctuation Annotationfrom Speech**. In *ICSLP*, Denver, Colorado, USA, September 2002. 40

[29] Q. Jin and T. Schultz. **Speaker Segmentation and Clustering in Meetings**. In *ICASSP*, Jeju Island, Korea, October 2004. 26

[30] S. Khadivi, A. Zolnay, and H. Ney. **Automatic Text Dictation in Computer-assisted Translation**. In *Interspeech*, Portugal, Lisbon, September 2005. 5

[31] P. Koehn. **European Parliament Proceedings Parallel Corpus 1996-2009**. Last accessed on February 4, 2010 at http://www.statmt.org/europarl/. 24

[32] P. Koehn. **A Parallel Corpus for Statistical Machine Translation**. In *MT Summit*, Phuket, Thailand, September 2005. 23, 32, 35, 79

[33] P. Koehn and C. Monz. **Manual and Automatic Evaluation of Machine Translation between European Languages**. In *Proc. on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, USA, 2006. 32, 86

[34] P. Koehn, F. J. Och, and D. Marcu. **Improved Backing-off for m-gram Language Modeling**. In *ICASSP*, Detroit, MI, USA, May 1995. 13

[35] P. Koehn, F. J. Och, and D. Marcu. **Statistical Phrase-Based Machine Translation**. In *HLT-NAACL*, Edmonton, Canada, April 2003. 17

[36] K. Kohn and S. Kalina. **The Strategic Dimension of Interpreting**. *Meta: Journal des traducteurs*, **41(1)**:118–138, 1996. 55

[37] L. Lamel, J.L. Gauvain, and G. Adda. **Investigating Lightly Supervised Acoustic Model Training**. In *ICASSP*, Salt Lake City, USA, May 2001. 6, 76, 104

[38] C.J. Leggetter and P.C. Woodland. **Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models**. *Computer Speech and Language*, pages 171–185, April 1995. 30

[39] Y. Liu. **Structural Event Detection for Rich Transcription of Speech**. PhD thesis, Purdue University, West Lafayette, IN, USA, 2004. 40

[40] L. Mangu, E. Brill, and A. Stolcke. **Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks**. *Computer Speech and Language*, pages 373–400, April 2000. 30, 79

[41] E. Matusov, S. Kanthak, and H. Ney. **On the Integration of Speech Recognition and Statistical Machine Translation**. In *Interspeech*, Portugal, Lisbon, September 2005. 18

[42] E. Matusov, G. Leusch, O. Bender, and H. Ney. **Evaluating Machine Translation Output with Automatic Sentence Segmentation.** In *IWSLT*, pages 148–154, Pittsburgh, USA, October 2005. 27, 45

[43] D. Mostefa, O. Hamon, N. Moreau, and K. Choukri. **TC-STAR Evaluation Report, Del.no 30**. www.tc-star.org, May 2007. 55, 59, 60

[44] H. Ney and S. Ortmanns. **Dynamic Programming Search for Continuous Speech Recognition**. *IEEE Signal Processing Magazine*, pages 64–83, September 1999. 12

[45] S. Novotney, R. Schwartz, and J. Ma. **Unsupervised Acoustic and Language Model Training with Small Amounts of Labelled Data**. In *ICASSP*, Taipei, Taiwan, March 2009. 7

[46] F. J. Och. **Minimum Error Rate Training in Statistical Machine Translation**. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160 – 167, Sapporo, Japan, 2003. 16, 34

REFERENCES

[47] F. J. Och and H. Ney. **Discriminative Training and Maximum Entropy Models for Statistical Machine Translation**. In *ACL*, Philadelphia, PA, USA, July 2002. 16

[48] F.J. Och and H. Ney. **Improved Statistical Alignment Models**. In *Proc. of ACL*, Hongkong, China, 2000. 32

[49] F.J. Och and H. Ney. **A Systematic Comparison of Various Statistical Alignment Models**. *Computational Linguistics*, **29(1)**:19–51, 2003. 86

[50] K. Papineni, S. Roukus ad T. Ward, and Wei-Jing Zhu. **BLEU: a Method for Automatic Evaluation of Machine Translation**. In *ACL*, Philadelphia, PA, USA, July 2002. 19, 20

[51] K. A. Papineni, S. Roukos, and R. T. Ward. **Maximum Likelihood and Discriminative Training of Direct Translation Models**. In *ICASSP*, Seattle, WA, USA, May 1998. 16

[52] M. Paulik, S. Rao, I. Lane, S. Vogel, and T. Schultz. **Sentence Segmentation and Punctuation Recovery for Spoken Language Translation**. In *ICASSP*, Las Vegas, NV, USA, April 2008. 39

[53] M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel. **Speech Translation Enhanced Automatic Speech Recognition**. In *Proc. of ASRU*, San Juan, Puerto Rico, 2005. 5, 58

[54] M. Paulik, S. Stüker, C. Fügen, T. Schultz, and A. Waibel. **Translating language with technology's help**. *IEEE potentials - the magazine for high-tech innovators*, **26(3)**:30–35, 2007. 5, 58, 72

[55] M. Paulik and A. Waibel. **Lightly Supervised Acoustic Model Training on EPPS Recordings**. In *Interspeech*, Brisbane, Australia, September 2008. 76

[56] M. Paulik and A. Waibel. **Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data**. In *ASRU*, Merano, Italy, December 2009. 85

[57] M. Paulik and A. Waibel. **Spoken Language Translation from Parallel Speech Audio: Simultaneous Interpretation as SLT Training Data**. In *ICASSP*, Dallas, TX, USA, March 2010. 101

[58] D. Povey. **Discriminative Training for Large Vocabulary Speech Recognition**. PhD thesis, University of Cambridge, Cambridge, UK, 2003. 14

[59] D. Povey, D. Kanevsky, B. Kingsbury, and B. Ramabhadran. **Boosted MMI for model and feature-space discriminative training**. In *ICASSP*, Las Vegas, LV, USA, April 2008. 30, 117

[60] J.R. Quinlan. **The C4.5 Induction System - C4.5 Release 8**. http://rulequest.com/Personal/. 45

[61] L. R. Rabiner. **A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition**. *Proceedings of the IEEE*, **77 (2)**:257–286, February 1989. 12

[62] I. Rogina. **Lecture and Presentation Tracking in an Intelligent Meeting Room**. In *International Conference on Multimodal Interfaces*, Pittsburgh, PA, USA, October 2002. 60

[63] K. Rottmann and S. Vogel. **Word Reordering in Statistical Machine Translation with a POS-based Distortion Model**. In *International Conference on Theoretical and Methodological Issues in MT*, Skövde, Sweden, September 2007. 32, 33

[64] Robert Santiago. **Consecutive Interpreting: A Brief Review**. accessed on January 11, 2010 at http://home.earthlink.net/∼terperto/id16.html, 2004. 54, 55

[65] R. Schlüter. **Investigations on Discriminative Training Criteria**. PhD thesis, Rheinisch Westflische Technische Hochschule, Aachen, Germany, 2000. 14

[66] T. Schultz and A. Waibel. **Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition**. *Speech Communication*, **35 (1-2)**:31–51, 2001. 7

[67] H. Schwenk. **Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation**. In *IWSLT*, Hawaii, USA, October 2008. 8

[68] D. Seleskovitch. **Interpreting for International Conferences**. *Pen and Booth, Arlington, VA*, 1978. 54

[69] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. **Prosody-based Automatic Segmentation of Speech into Sentences and Topics**. *Speech Communication*, pages 127–154, 2000. 40

[70] F. Smith. *Reading Without Nonsense*. NY Teachers College Press, NY, NY, 1985. 54

[71] H. Soltau, F. Metze, C. Fügen, and A. Waibel. **A One Pass-decoder Based on Polymorphic Linguistic Context Assignment**. In *ASRU*, Madonna di Campiglio Trento, Italy, December 2001. 14, 29

[72] A. Stolcke. **SRILM – an extensible language modeling toolkit**. In *Intl. Conf. on Spoken Language Processing*, Denver, CO, USA, September 2002. 29

[73] S. Stüker. **Acoustic Modelling for Under-Resourced Languages**. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2009. 125

[74] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y. Tam, and M. Wölfel. **The ISL TC-STAR Spring 2006 ASR Evalutation Systems**. In *TC-STAR Workshop*, Barcelona, Spain, 2006. 30

[75] J. Tiedemann. **Combining Clues for Word Alignment**. In *EACL*, Budapest, Hungary, April 2003. 69

[76] N. Ueffing. **Using Monolingual Source-Language Data to Improve MT Performance**. In *IWSLT*, Kyoto, Japan, November 2006. 7

[77] European Union. **Frequently Asked Questions about the European Union's policy on languages**. Last accessed on January 29, 2010 at http://europa.eu/languages/en/document/59l. 21

[78] S. Vogel. **SMT Decoder Dissected: Word Reordering**. In *Proc. of Coling*, Beijing, China, 2003. 17, 33

[79] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. **Meeting Browser: Tracking and Summarizing Meetings**. In *DARPA Broadcast News Workshop*, 1998. 60

[80] A. Waibel and C. Fügen. **Spoken Language Translation – Enabling Cross-Lingual Human-Human Communication**. *IEEE Signal Processing Magazine*, **25 (3)**:70–79, May 2008. 19

[81] M.C. Wölfel and J.W. McDonough. **Minimum Variance Distortionless Response Spectralestimation, review and refinements**. In *IEEE Signal Processing Magazine*, pages 117–126, 2005. 29, 79

[82] P. Zhan and M. Westphal. **Speaker Normalization Based On Frequency Warping**. In *ICASSP*, Munich, Germany, April 1997. 30

[83] R. Zhang, G. Kikui, H. Yamamoto, T. Watanabe, F. Soong, and W. K. Lo. **A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation**. In *COLING*, Geneva, Switzerland, August 2004. 18

[84] Y. Zhang and S. Vogel. **Suffix Array and its Applications in Empirical Natural Language Processing**. In *the Technical Report CMU-LTI-06-010*, Pittsburgh, USA, December 2006. 34

[85] X. Zhu and R. Rosenfeld. **Improving Trigram Language Modeling with the World Wide Web**. In *ICASSP*, Salt Lake City, Utah, USA, March 2001. 7