Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 27 (2013) 455–474
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Training speech translation from audio recordings of interpreter-mediated communication☆

Matthias Paulik[a,*,1], Alex Waibel[b,c]

[a] *Cisco Systems, 170 W Tasman Dr, San Jose, CA 95134, USA*
[b] *Carnegie Mellon University, USA*
[c] *Karlsruhe Institute of Technology, Germany*

Received 12 February 2010; received in revised form 11 February 2011; accepted 15 April 2011
Available online 10 May 2011

Abstract

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 speech translation. However, the development of deployable speech translation systems still happens only for a handful of languages. Prohibitively high costs attached to the acquisition of sufficient amounts of suitable speech translation training data are one of the main reasons for this situation. A new language pair or domain is typically only considered for speech translation development after a major need for cross-lingual verbal communication just arose—justifying the high development costs. In such situations, communication has to rely on the help of interpreters, while massive data collections for system development are conducted in parallel. We propose an alternative to this time-consuming and costly parallel effort. By training speech translation directly on audio recordings of interpreter-mediated communication, we omit most of the manual transcription effort and all of the manual translation effort that characterizes traditional speech translation development.

Keywords: Speech translation; Spoken language translation; Parallel speech

1. Introduction

Speech translation (ST) 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 and constitutes as such one of the major research areas of speech and language processing. The impressive advances made in ST can mostly be attributed to the statistical modeling schemes that nowadays dominate its component technologies, automatic speech recognition (ASR) and machine translation (MT). However, the performance of these statistical models depends heavily on an effective model parameter estimation from vast amounts of training data. Considering the statistical

☆ This paper has been recommended for acceptance by 'Shrikanth Narayanan'.
* Corresponding author. Tel.: +1 408 853 8437.
E-mail addresses: mapaulik@cisco.com, matthias.paulik@gmail.com (M. Paulik).
[1] This work was done while the author was working at Carnegie Mellon University and Karlsruhe Institute of Technology.

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.04.001

models involved in state-of-the-art ST, the translation model (TM) can be identified as the most demanding in terms of training data requirements. While domain specific monolingual resources—transcribed speech audio for acoustic model (AM) training and text corpora for language model (LM) training—can be hard to acquire, domain-specific bi-lingual text data of sentence-aligned manual translations for TM training are even more scarce and expensive to create. The high cost attached to acquiring sufficient amounts of suitable ST training data is one of the main reasons why the development of deployable ST systems still happens only for a handful of economically or politically viable languages. Speech translation development for a new language pair or domain typically occurs only after a major need for cross-lingual verbal communication just arose. In such situations, communication has to be achieved with the help of interpreters, while massive ST training data collections are conducted in parallel. We propose an alternative to this time-consuming and costly parallel effort. By training ST directly on audio recordings of interpreter-mediated communication scenarios, we omit most of the manual transcription effort and all of the manual translation effort that characterizes traditional speech translation development. Although parallel speech (pSp)—interpreter speech together with the source language speech, from which the interpretation is being rendered—is fundamentally different from parallel text, our experiments show that translation models (TMs) can be successfully trained at surprisingly high performance levels from pSp audio alone. We thus eliminate the need for parallel text data for TM training. We further minimize the amount of costly supervision necessary for ST training by applying unsupervised and pSp-supervised AM and LM training techniques within a framework that automatically extracts ST training data from pSp audio

Learning automatic *translation* (of speech or text) from *interpretation* audio faces major challenges, many of which stem from the nature of interpretation. In Section 2 we take a closer look at interpretation and compare it to translation. Section 3 introduces our approach for training statistical TMs from pSp audio of simultaneous interpretation (SI). We use English/Spanish (En/Sp) SI as provided during European Parliament Plenary Sessions (EPPS). Section 4 introduces our framework for exploiting pSp audio in a fully automatic, unsupervised manner as a data resource for training all major models involved in state-of-the-art ST. In Section 5 we transfer our most important findings to the pSp audio of consecutive interpretation (CI) between the resource-rich language English and the resource-deficient language Pashto. Finally, in Section 6 we summarize our results, discuss their significance for speech translation development and examine several remaining challenges.

2. Translation, interpretation and parallel speech

2.1. Terminology

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

2.2. The nature of interpretation

Fig. 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 employed by interpreters.

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 (Seleskovitch, 1978). 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 (Santiago, 2010). The reason for this lies within the limitations of 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 (Smith, 1985).

Simultaneous Interpretation	Consecutive Interpretation
SPANISH UTTERANCE: "trataremos de que todo el personal tenga"	ENGLISH UTTERANCE: okay and what is the importance of this gas station
TRANSLATION: "we shall try that all the staff will get"	PARALLEL SPEECH: بەيخي صحيح دە د دي پە ەكلە تا خە غوښتل چې زما سر ووايي
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 ..."	*TRANSLATION: it is okay - what do you want to tell me about this*

Fig. 1. Interpretation (parallel speech) versus translation.

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 (Besien, 1999) and compensatory strategies (Al-Kahnjii et al., 2000). 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 (Mostefa et al., 2007) suggest that 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, in Mostefa et al. (2007), a significantly higher *effective* information loss of 29% was reported. The effective information loss was measured by counting the amount of comprehension questions that were not correctly answered by human evaluators allowed to listen to the recorded interpreter speech twice.[2] This effective information loss results from a combination of missing information (9%) and the difficulty of human evaluators to follow the flow of interpreter speech, which is often syntactically ill-formed.

In CI, interpreters face less severe time constraints, resulting in more accurate, equivalent, and complete interpretations (Santiago, 2010). 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" (Kohn and Kalina, 1996).

2.3. Interpretation and (automatic) translation

To measure the quality of automatic translation of speech or text we apply the widely adopted BLEU metric (Papineni et al., 2002), which, on a scale from 0 to 100, compares MT output to one or more human reference translations based on *n*-gram comparison. Like all machine evaluations, BLEU is not able to determine if a given MT output correctly translated all meaning, but can only determine how 'similar' (in terms of *n*-gram matches) the output is to given reference translations. As described in detail in Section 2.2, 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 EPPS development set 'TDev' (compare Section 3 for a detailed description of TDev). Given two reference translations, only 14.2 BLEU for Sp → En and 18.2 BLEU En → Sp are achieved. This compares to BLEU scores above 40, achieved by state-of-the art ST systems (trained on approximately 100 h of transcribed speech and 30+ million translated words), as for example developed within the European TC-STAR project.

[2] The evaluators were allowed to interrupt the playback to write down their answers.

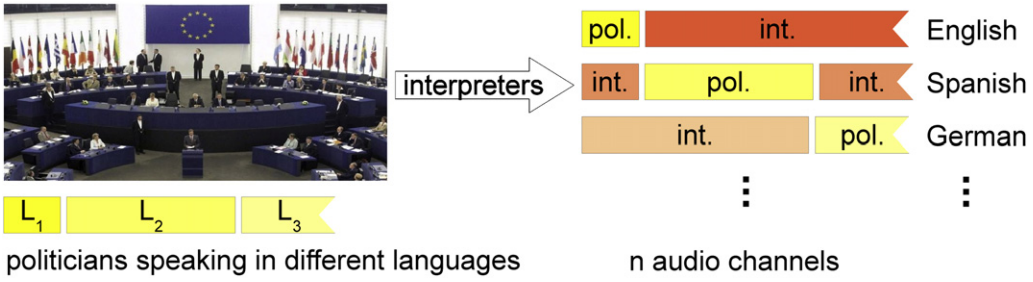

Fig. 2. EPPS live broadcast: a data source for parallel speech audio.

3. Automatic translation from simultaneous interpretation

In this section we introduce our approach for training statistical TMs from pSp audio, as it was first introduced by us in Paulik and Waibel (2009). Our experiments are based on pSp audio from En/Sp SI, as provided during European Parliament Plenary Sessions. As motivated in the introduction, training TMs from pSp audio is of special interest for ST development between new language pairs where pre-existing data resources for traditional ST 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.

3.1. General approach and major challenges

We use pSp audio within a standard training setup for phrase-based statistical MT. To do so, we transcribe the pSp using ASR. The resulting ASR hypotheses are aligned on a speech utterance basis by applying special alignment strategies that are tailored to the pSp of SI. 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. The alignment procedure is described in detail in Section 3.4. Our standard training setup extracts phrase tables from the created bilingual training corpus by using the GIZA++ toolkit (Och and Ney, 2003), in combination with University Edinburgh's training scripts, as provided during the NAACL 2006 Workshop on Statistical Machine Translation (Koehn et al., 2006). The GIZA++ toolkit is run with its standard parameter settings.

The major challenges faced include (a) the significant difference between translation and interpretation, (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).

3.2. Data and scoring

European Parliament Plenary Sessions are broadcast live via satellite in the different official languages of the European Union. Each language L_i has a dedicated audio channel. An interpreter provides the SI 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. The language-dedicated audio channels, depicted on the right hand side of Fig. 2, form an excellent source of pSp audio. In addition to the live broadcasts, the proceedings of the parliament are also published in the form of so-called final text editions (FTEs) in all official languages after approximately two months. These FTEs are created by a multitude of human transcribers and translators from recordings of the original politician's speeches.

For our experiments, we use the EPPS part of the English and Spanish TC-STAR spring 2007 verbatim task development and evaluation sets; in the following they are simply referred to as 'dev' and 'eval'. It has to be noted that these sets only consist of politician speech and do not include any interpreter speech. This stands in contrast to our pSp corpus. This corpus was collected in-house by recording the live broadcast English and Spanish audio channels, which both contain a mix of politician and interpreter speech. AM training data, as provided during the TC-STAR evaluation campaign, also consists of a mix of politician and interpreter speech. Our language models are (mostly) estimated on

Table 1
Data statistics: parallel speech, dev and eval sets.

	English			Spanish		
	pSp	dev	eval	pSp	dev	eval
Speech utt.	65.3k	1287	1926	63.2k	1707	2085
Transl utt.	65.3k	1194	1167	63.2k	792	746
Words [k]	954.4	27.9	26.0	897.0	22.4	25.8
Audio [h]	111.3	3.2	2.7	105.2	2.3	2.7

the FTEs or, in the case of the constrained Spanish ASR (compare Section 3.3), exclusively on the human reference transcription as used for acoustic model training. Detailed data statistics for dev, eval and the pSp corpus are given in Table 1. For dev and eval, the number of speech utterances differs from the number of translation utterances due to the mismatch of automatic speech/non-speech segmentation applied prior to ASR and the manual utterance segmentation of the reference translations. The number of words in the pSp corpus are estimated on the first-best hypotheses of our standard Spanish and English ASR systems, as no reference transcription of the pSp corpus is available. The pSp corpus is comprised of sessions from the time periods 04 May 2005–26 May 2005 and 08 September 2005–01 June 2006. The development set has sessions from 06 June 2005 to 09 September 2005 and the evaluation set has sessions from 12 June 2006 to 28 September 2006. AM training data is from the time period May 2004 to June 2005. The FTEs are from the time period April 1996 to May 2005.

In addition to our development and evaluation set, we use one additional Parliamentary session from 26 October 2004 to tune our alignment algorithm presented in Section 3.4. We refer to this session in the following as 'TDev'. We extracted this from the TC-STAR verbatim 2005 development set. In contrast to dev and eval, TDev is (a) comprised of a mix of politician and interpreter speech (b) forms a parallel speech corpus and (c) is provided with a manual utterance segmentation that is kept consistent for the reference transcriptions and verbatim reference translations. This means in particular, that for all English and Spanish speech utterances, properly aligned transcription references and translation references are available for TDev. The English (Spanish) part TDev consists of 1256 (1589) utterances, with 17.4k (14.7k) running words and 95 (89) minutes of audio.

For scoring ASR and MT performance we use lower-cased references without punctuation. ASR performance is measured in word error rate (WER) and MT performance is measured in BLEU using two reference translations. We use the multiple reference word error rate (mWER) segmentation script (Matusov et al., 2005), as it was provided by RWTH Aachen University within TC-STAR, to align the translated speech utterances to the translation references.

3.3. ASR and MT systems

The employed ASR systems are developed with the Janus Recognition Toolkit (JRTk). The SRI Language Model Toolkit (Stolcke, 2002) is used for LM training. Table 2 gives an overview of the WER and LM perplexity (PPL) of the English and (unconstrained) Spanish ASR.

Both, English and (unconstrained) Spanish ASR, feature two decoding passes and two ASR sub-systems per pass, with unsupervised speaker adaptation in the second pass. At the end of each decoding pass, the output of the individual ASR sub-system is combined via confusion network combination (Mangu et al., 2000). The systems are trained on 80+ hours of transcribed speech data and 70+ million words of monolingual text data. For a detailed system description, refer to Paulik and Waibel (2009). In addition to the standard Spanish ASR system we use two constrained Spanish

Table 2
English and (unconstrained) Spanish ASR: language model perplexity (PPL) and word error rate on dev and eval.

	dev		eval	
	Spanish	English	Spanish	English
PPL	89	108	89	106
WER	8.4	13.9	9.0	12.2

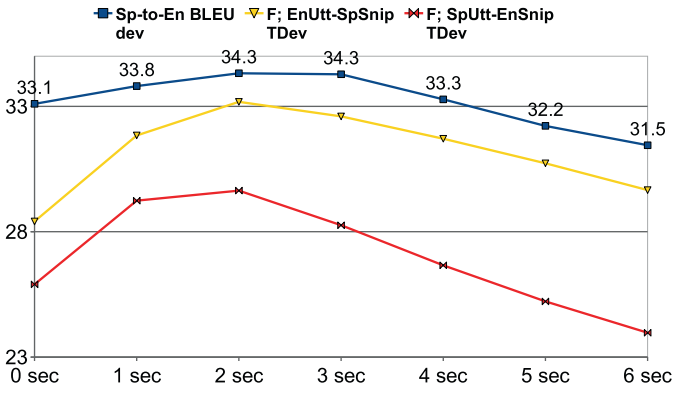

Fig. 3. F-measure on TDev and Sp → En BLEU score on dev for different target speech snippet padding values $x \in \{0, 1, \ldots, 5, 6\}$ s.

ASR systems 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 a weak AM. Both contribute to an increased WER. To simulate resource limitation, we first constrain the Spanish LM to the 748k running words of the transcriptions that were used to train the AM. The constrained LM yields a perplexity (PPL) of 178 on dev and a PPL of 177 on eval, resulting in word error rates of 16.1% and 16.5%, respectively. To simulate a lower performance AM we further limit the system to a context independent phone-set. Constraining the AM results in a WER of 33.3% on dev and 33.1% on eval.

In the following, we refer to the three described Spanish ASR systems as **Sp-ASR**$_9$, **Sp-ASR**$_{16}$, **Sp-ASR**$_{33}$. This naming scheme reflects the WER of these systems on our development and evaluation sets.

For MT, we use the ISL beam search decoder (Vogel, 2003). The decoder includes a simple word-reordering model that assigns higher costs to longer distance reordering. We use a reordering window of 4. The applied target LMs are identical to the LMs used for ASR. To optimize the system towards a maximal BLEU score, we apply Minimum Error Rate (MER) Training (Och, 2003).

3.4. Aligning SI speech

If 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 dev and eval, 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. Also, speakers change frequently in both channels. 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 TDev and for different values of x, we compute the F-measure (harmonic mean of precision and recall) in respect to uni-gram matches between the padded, automatically transcribed pSp snippets and the TDev available reference translations. Fig. 3 depicts how the F-measure changes for different values of x. It shows a peak at $x = 2$ s 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 $x \in \{0, 1, \ldots, 5, 6\}$ s. After extracting seven different phrase tables from these MT training corpora, we compute the translation performance for Sp → En on dev, using these phrase tables. As we can see in

SPANISH UTT @ TIME 1392.746:
"hay que terminar los las negociaciones"

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

PARALLEL SPEECH SNIPPET:
"so we have our buildings policy for brussels there are two [TIME: 1388.497] buildings in particular that account for more than three hundred million euros [TIME: 1391.91] the two buildings there in brussels **we have negotiations under way they shall be concluded** [TIME: 1396.581] i hope before the end of the year [TIME: 1399.516] and that would mean that we could already start making some of the payments in the year two thousand and five"

Fig. 4. ±6 s utterance based padding of parallel speech.

Fig. 3, the BLEU score again peaks at $x = 2$s, showing that the F-measure computed on TDev correlates well with the translation performance on dev. In other words, the optimal padding value x for aligning our pSp corpus can be well predicted by simply computing the F-measure on TDev.

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 TDev, we find that, if the information contained in the source utterance is at all present in the parallel speech, a 6 s utterance padding almost always guarantees that the information can be found in the respective target audio snippet. By a 6 s 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 s. Fig. 4 gives an example of parallel speech that is aligned based on a 6 s 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 s 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. The algorithms seeks to optimally cut down the 6 s padded target speech snippet of a source utterance z by first finding the optimal left cut position and then finding the optimal right cut position. This is accomplished with the help of alignment weights a introduced between source and target words. The left cut position c_i defines all target words $w_1, \ldots, w_i$ as not being aligned to any source word in z, and all target words $w_{i+1}, \ldots, w_m$ as not being aligned to any word found in neighboring source utterances $u \neq z$. The optimal left cut position is found by maximizing the sum over all alignment weights:

$$arg\max_i \sum_{n=0}^{i} a_n^{u \neq z} + \sum_{n=i+1}^{m} a_n^{u=z} \tag{1}$$

The right cut position is found accordingly. The alignment weights associated with source and target word pairs are based on forward and backward IBM4 word-to-word translation probabilities. The IBM4 lexicons are estimated via GIZA++ training on the parallel speech corpus resulting from our simple 2 s word-time based alignment scheme. Alignment weights between a source and target word pair are greater than zero if the combined forward and backward word 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. Further, the alignment weight between a 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:

$$importance(tw) = 1.0 - sL * LM(tw) \tag{2}$$

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. During computation of the optimal left and right cut positions and in addition to the introduced alignment weights a, we further consider alignment weight clusters formed by target language bi- and tri-grams. For

Table 3
Influence of the 2-pass alignment strategy on the Sp → En MT performance. The table lists the BLEU scores for using Spanish reference transcriptions (0% WER) as input to MT and using TMs trained on parallel speech transcribed with our three Spanish ASR systems.

	dev			eval		
Sp-ASR$_X$	9	16	33	9	16	33
±2 s	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

each such cluster we introduce additional alignment weights that are included in the overall sum. The alignment weight *aBI* for a bi-gram alignment cluster formed by the alignment weights a_1 and a_2 is, for example, given as:

$$aBI(a_1, a_2) = (a_1 * a_2)^{bw} \tag{3}$$

with the bi-gram weight *bw* to allow for a flexible additional weighting of such bi-gram alignment weight clusters.

To optimize the described parameters of our two-pass alignment algorithm, we perform a grid search on TDev, aiming for a maximal value of F-measure. The F-measure is based on matching uni-grams in the pSp snippets and the reference translations. Table 3 lists the Sp → En MT performance when using the two described alignment strategies and automatically transcribing the pSp corpus at different Spanish WER levels (9, 16 and 33%). At all three Spanish WER 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 MT performance degrades only by approximately 12% relative on the eval set. This indicates that training TMs from automatically transcribed pSp is robust to strong variations in ASR performance on one side of the pSp corpus.

3.5. Machine translation and speech translation results

Table 4 lists the Sp → En and En → Sp MT results obtained when using TMs trained from pSp. We list the results for pSp that was transcribed at the different Spanish WER levels provided by our three Spanish ASR systems (approximately 9, 16 and 33% WER). The English ASR system was kept unchanged; we estimate its WER at approximately 12–14% WER, given its performance on dev and eval. BLEU scores marked with $_c$ were computed using the constrained Spanish LM. For comparison, we also list results when training TMs 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 sentences pairs from the part of the TC-STAR training corpus that corresponds to the same time period as the parallel speech, 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 FTEs. 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 includes punctuation removal, expansion of abbreviations and conversion of numbers and dates to their spoken form.

Table 4
Training corpus dependent MT performance in BLEU. Results are shown for using a translation model training corpus of manual translations (first data row) and pSp training corpora, obtained with our three different Spanish ASR systems.

Training corp.	dev		eval	
Type, Sp-ASR$_x$	Sp-En	En-Sp	Sp-En	En-Sp
Translations, n/a	44.5	48.0	43.9	40.9
pSp, Sp-ASR$_9$	35.1	39.7	34.3	31.2
pSp, Sp-ASR$_{16}$	33.5	34.5_c	33.5	27.0_c
pSp, Sp-ASR$_{33}$	29.1	31.1_c	30.1	23.3_c

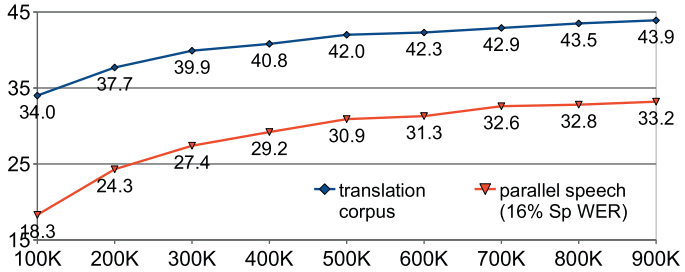

Fig. 5. BLEU score (*y*-axis) dependent on training corpus type (parallel speech vs. manual translation) and training corpus size in steps of 100k running words.

The results show degraded translation performance for training TMs 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 3.4, 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.

We also compute the MT performance on TDev for our best pSp-trained models. We achieve BLEU scores of 43.5 and 34.8 for Spa → Eng and Eng → Spa, respectively. These scores are 2–3 times higher than the BLEU scores we estimated for the TDev manual transcription of parallel interpreter speech; compare also Section 2.3. In other words, our translation models, despite being trained solely on (large amounts) of pSp, provide an automatic translation of TDev that is significantly more similar (in terms of *n*-gram matches) to the manual translation references than the pSp of TDev itself is.

The highest achieved Sp → En translation performance of 34.3 BLEU is on the same level as the translation performance of TMs 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, TMs trained on pSp audio of En/Sp SI require 10 times more data (measured in number of translated/interpreted words) than TMs trained on manual translations, to reach similar BLEU performance. Fig. 5 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 TMs 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 10.5–11 points (a 24.0–26.4% 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 5 lists the speech translation results on eval. The WER 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 use the same decoder

Table 5
Training corpus dependent speech translation performance in BLEU. The WER of the ASR first-best hypotheses used as input to the MT system are listed in the first row. Translation model training is either based on a training corpus of manual translation, or on a training corpus of automatically transcribed parallel speech.

Training corp.	Sp-En			En-Sp
Type, Sp-ASR$_x$	9.0%	16.5%	33.1%	12.2%
Translations, n/a	40.0	36.0	27.8	33.8
pSp, Sp-ASR$_9$	31.5	–	–	26.1
pSp, Sp-ASR$_{16}$	–	29.1	–	22.8$_c$
pSp, Sp-ASR$_{33}$	–	–	21.0	19.8$_c$

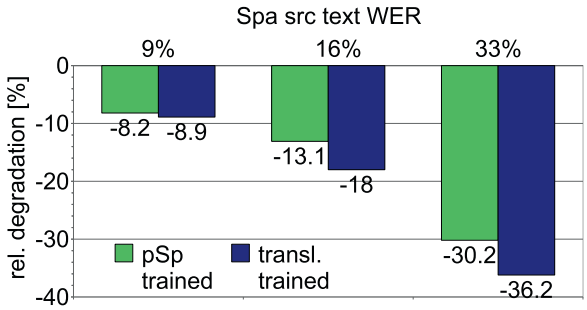

Fig. 6. Relative degradation in BLEU (Sp → En), dependent on Spanish input WER and training corpus type (pSp vs. manual translations).

weights found via MER optimization on the dev 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 dev. Compared to TMs trained on a similarly sized bilingual text corpus of translations, we observe a degradation of approximately 8 BLEU points when using parallel speech trained TMs. This degradation in performance is almost 2 BLEU points less than in the case of MT. In general, the relative degradation in BLEU caused by recognition errors in the input text is smaller for pSp-trained TMs, as depicted in Fig. 6. 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 eval for speech translation. The smaller degradation in BLEU score indicates that the pSp-trained TMs 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 ST performance experienced by using SI instead of translation for TM training.

4. Parallel speech audio as ST training data

So far we demonstrated that statistical TMs can be trained in a fully automatic, unsupervised manner from audio recordings of SI. In this section we extend the use of pSp audio as a data source for unsupervised training of all major models involved in statistical ST: the ASR acoustic model and ASR language model as well as the MT translation model and MT target LM (Paulik and Waibel, 2010). Specifically, we explore techniques for unsupervised AM and LM training. Further, we exploit the parallel nature of pSp audio to train TMs and to introduce light supervision for model training. We conduct our experiments on a subset of the En/Sp pSp corpus used in Section 3. To simulate the setting of ST between a resource rich and a resource-deficient language, we limit the supervised training data for the Spanish models to 10 h of manually transcribed Spanish audio and to a parallel text corpus comprised of 100k translated words. We also consider the situation where no parallel text data is available. We seek to improve the performance of the statistical models affected by resource deficiency: the Spanish AM, Spanish LM and En ↔ Sp TMs.

4.1. System architecture

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 extracting training data from pSp audio. Fig. 7 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 extract ST training data 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. Similar to previous works (Khadivi et al., 2005; Reddy and Rose, 2008; Paulik and Waibel, 2008b), we exploit the parallel

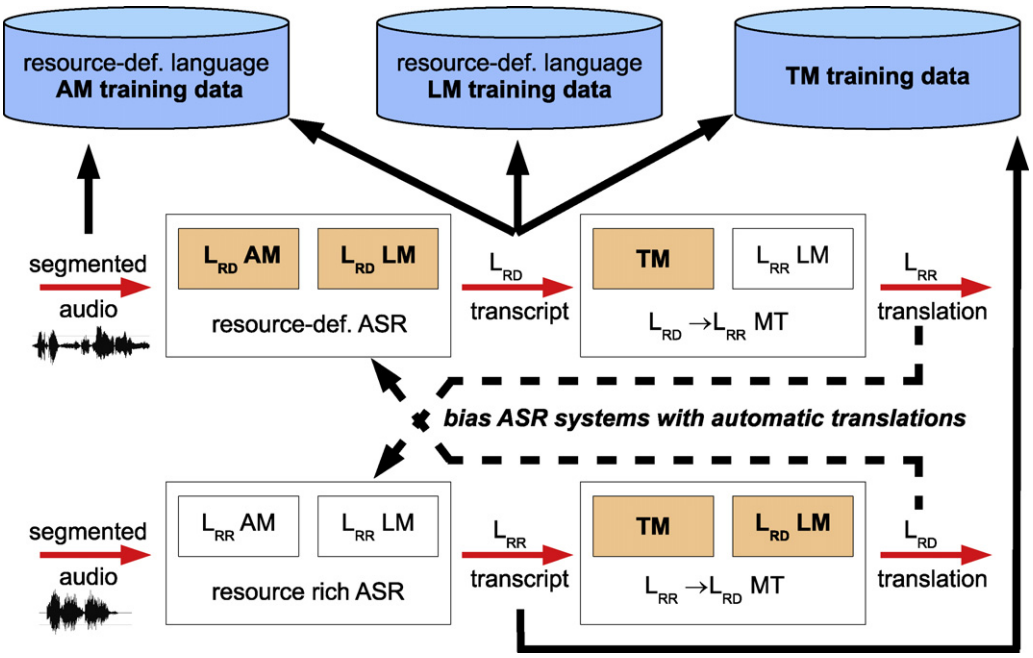

Fig. 7. Extracting speech translation training data from parallel speech.

information given in the respective other language audio stream to bias the ASR systems for improved transcription performance. In the proposed context, such improved ASR performance directly affects the quality of the extracted training data.

4.2. Data and baseline systems

We use the same En/Sp development and evaluation sets as described in Section 3; TDev, dev and eval. The pSp audio corpus is a subset of the pSp audio corpus from Section 3, consisting of 67 sessions from the time period 08 September 2005 to 01 June 2006. The supervised Spanish ST training data is limited to 10 h of manually transcribed Spanish audio and to a parallel text corpus comprised of 100k translated words. Detailed data statistics for the training sets are shown in Table 6.

The MT decoder, MT training procedure and English ASR are identical to the ones described in Section 3. For training TMs from pSp audio of SI, we rely on the more simple utterance alignment strategy of padding the target parallel speech snippets with 2 s. 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 10 h Spanish EPPS data via three iterations of Viterbi training. The 3-gram LM is estimated on 179.6k running words from the AM training data reference 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 7.

Table 6
Data statistics: Spanish ST training data.

	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 7
English and Spanish baseline word error rates.

	English-to-Spanish			Spanish-to-English		
	TDev	dev	eval	TDev	dev	eval
Base WER	13.1	13.9	12.2	26.1	26.9	27.1

4.3. PSP audio for ASR model training

Unsupervised acoustic model training (Kemp and Waibel, 1999; Lamel and Gauvain, 2002) 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 (Lamel et al., 2001) 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 proposed in Paulik and Waibel (2008a).

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 build 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 s 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 TDev via a grid search. Table 8 shows the influence of the session LM bias and the combination of utterance LM bias and session LM bias on the Spanish WER on TDev; the WER is reduced by 6% relative from 26.1% to 24.5%.

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 10 h of manually transcribed audio plus 92 h of automatically transcribed audio. Results obtained with the re-trained AMs are listed in Table 9, along with results for unsupervised LM training. The first two columns of Table 9 specify if the baseline AM/LM was used or a model trained with the additional 92 h 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 TDev, when biasing the ASR with knowledge from the English parallel speech. Since we do

Table 8
Biasing ASR with pSp; Spanish WER on TDev.

Baseline	Session bias	Session & utt. bias
26.1	25.4	24.5

Table 9
Re-raining the Spanish acoustic model and language model with additional 92 h 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.

AM	LM	TDev	dev	eval
Base	Base	26.1 [24.5_b]	26.9	27.1
+92 h	Base	24.0 [23.0_b]	24.9	25.5
Base	+92 h	24.5 [23.3_b]	25.7	25.5
+92 h	+92 h	22.5 [21.5_b]	24.0	24.2
+92 h_b	+92 h_b	22.0 [21.6_b]	23.5	23.8

Table 10
LM re-training with additional 92 h of automatically transcribed Spanish parallel speech: influence on perplexity.

LM	TDev	dev	eval
Base	182	269	276
+92 h	129*	202	206
+92 h_b	127*	200	204

not have English pSp available for dev and eval, such a bias is not possible on these data sets. The results show that light supervision during training benefits ASR performance.

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 10 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 TDev PPLs (marked by*) does not include automatic transcriptions of TDev 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 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 do not include ASR transcripts of the very same session.

4.4. PSP audio for MT model training

When traditional MT training data is available, it is necessary to determine how traditionally trained TMs can benefit most from additional pSp audio. By simply adding our pSp training data to the 100k translated words parallel text corpus, we observe only small improvements. 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 to the pSp part. We achieve this higher weighting by simply duplicating the supervised training corpus x times. Fig. 8 gives an overview of the Sp → En text translation (0% WER of the Spanish input) results on the dev set for $x \in \{1, 2, \ldots, 7, 8\}$. 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

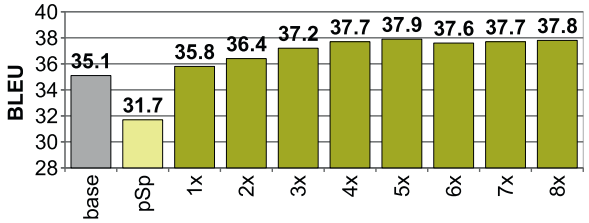

Fig. 8. Combining parallel speech (pSp) training data with our baseline parallel text training corpus. The baseline training corpus of manual translation receives a higher weight by repeating it x times. Results are shown for Sp → En text translation on dev.

Table 11
Translation model re-training with additional 92 h of automatically transcribed Spanish parallel speech: Sp → En text translation results in BLEU.

TM	TDev	dev	eval
Base	41.1	35.1	35.2
+92 h	44.5	37.9	37.8

Table 12
TM training: En → Sp text translation, BLEU score.

LM	TM	TDev	dev	eval
Base	Base	26.0	27.2	25.7
+92 h	Base	27.9	29.1	27.5
Base	+92 h	27.7	28.3	27.6
+92 h	+92 h	30.5	30.6	29.6

Table 13
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).

LM_{MT}	TM	TDev	dev	eval
Base	Base	24.0	22.4	21.6
+92 h	+92 h	28.5	25.7	25.2
+92 h	92 h	23.8	20.4	19.9

best translation results are achieved by adding the supervised parallel text corpus of manual translations 5 times to the combined training corpus. Tables 11 and 12 list the achieved text translation results for both translation directions. The presented results were obtained with ASR transcriptions created with the baseline ASR systems.

4.5. 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 10 h of transcribed Spanish AM data is available, but no baseline MT.

Table 13 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 92 h of pSp audio, we achieve a translation performance of 19.9 BLEU points. In Table 14 we show speech translation results for Sp → En.

Table 14
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 pSp 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 TM training. All other results are based on pSp training data transcribed with the Spanish baseline ASR system.

AM	LM_{ASR}	TM	TDev	dev	eval
Base	Base	Base	31.2	25.1	25.3
+92 h	+92 h	+92 h	34.8	28.0	28.3
+92 h_b	+92 h_b	+92 h_b	35.7	28.8	28.4
+92 h	+92 h	92 h	31.8	24.2	24.9
+92 h	+92 h	92 $h_{i=1}$	32.7	25.2	25.6

Table 15
Parallel speech audio statistics.

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

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 4.3. 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 again with transcriptions from systems that already benefit from re-trained models.

5. Speech translation from consecutive interpretation

In the previous sections we explored pSp audio of En/Sp SI as a resource for training ST systems. We identified parallel text data as one of the ST training resources that is especially hard to acquire. Therefore, we consider the result that TMs can (a) be trained from scratch with pSp audio, and (b) be improved with pSp audio as additional training data, as one of the most important findings. Our experiments suggest that pSp-trained TMs mirror the training corpus size-dependent performance of parallel text trained TMs, just at a lower level. We estimate the 'yield' of En/Sp SI audio to be around $n \cdot 10^{-1}$, meaning that we can expect 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. The yield of pSp audio certainly depends on different factors, as for example the type and 'quality' of used interpretation (CI vs. SI), 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 En/Sp 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-to-English (Pa $\rightarrow$ En) speech translation on the basis of pSp audio from consecutive interpretation. Specifically, we explore English/Pashto (En/Pa) pSp audio as (a) the sole data source for TM training, and (b) as additional training data, that is to be mixed with parallel text.

5.1. Experimental setup

Our experiments are based on data resources provided within US Darpa's TransTac project. TransTac aims to rapidly develop ST for real-word tactical situations. Typical scenarios are in the form of interviews, where an English-speaking soldier interviews a Pashto-speaking Afghani. Only very limited amounts of data resources are available for En/Pa ST development. Table 15 lists the statistics of the En/Pa 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. The CI example depicted at the very beginning, in Fig. 1, is an excerpt from this corpus. In contrast to our En/Sp 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' En/Pa parallel text corpus of manual

Table 16
Pa → En development and test set.

	Pa → En	
	Dev	Eval
Audio [min]	45.8	24.0
Words [k]	6.7	3.6

Table 17
Parallel speech audio: LM perplexity and ASR word error rate.

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

translations.[3] This parallel text corpus has 12.4k translated Pashto respondent utterances. The Pashto part comprises 260k words; the English part has 214k words. Table 16 lists the statistics of the Pa → En development and test sets. Both sets are based on native Pashto respondent speech and feature only one reference translation for scoring.

In order to utilize the En/Pa 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.

5.2. CI as only data source

In a situation where only untranscribed pSp audio of CI is available, the minimal requirement for ST 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, 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 Pa → En speech translation and to be able to automatically transcribe additional pSp audio, we train a Pashto ASR system on the 25.2 h 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 17 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 Pa → En development and test set is 33.7% and 33.9%, respectively.[4] 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 pSp corpus. System A uses TMs trained on the manually transcribed and translated Pashto respondent speech that is present in the

[3] This corpus corresponds to manually translated TransTac monolingual data collection corpus.

[4] English and Pashto ASR both feature only one decoding pass and small models tuned to the real-time requirements of our TransTac online evaluation system.

Table 18
Pa → En text and speech translation performance for systems A, B and C. The Pashto part of the parallel TM 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% WER) interpreter speech.

	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

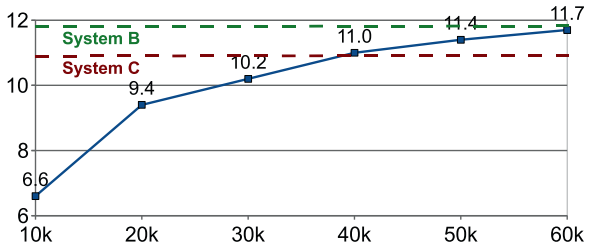

Fig. 9. Corpus-size dependent Pa → En BLEU score on dev of system A (trained on manual translation).

pSp 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 WER is on the same level as the worst WER level we considered En/Sp pSp audio. Table 18 lists the text and speech translation performance in BLEU for all three systems. 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 Fig. 9. It shows that our prediction was accurate. Another important hypothesis is that the translation performance of pSp trained TMs mirrors the translation performance of parallel text trained TMs, just at a lower level. While only a very limited amount of En/Pa pSp audio is available, we still can observe the same trend as observed for En/Sp pSp audio, when we compute the corpus-size dependent text translation performance of system A and B in increments of 90k words, compare Fig. 10.

5.3. CI and parallel text

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 TMs, resulting in systems D + A, D + B and D + C. Table 19 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

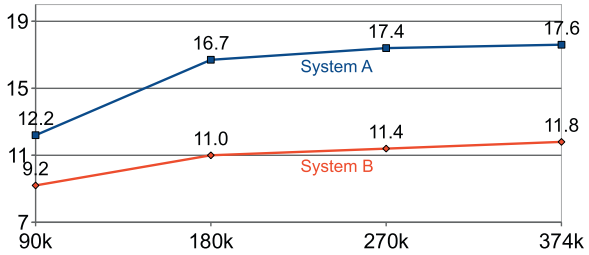

Fig. 10. Corpus-size dependent Pa → En BLEU scores on dev of system A (trained on manual translation) and B (trained on manually transcribed interpretation).

Table 19
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.

	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**

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 row of Table 19. 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 4.3, 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.

6. Results and discussion

We have explored parallel speech audio, as it occurs in interpreter-mediated communication, as a resource for training speech translation systems. Specifically, we have shown how training data can be extracted in a fully automatic, unsupervised manner from pSp audio and how this training data can be successfully used for the development of statistical speech translation. We applied techniques for unsupervised acoustic model and language model training. Further, we exploited the parallel nature of pSp audio to train translation models from audio and to introduce light supervision. We concentrate on scenarios that involve ST between a resource-rich and a resource-deficient language, as we argue that the additional data resource of parallel speech audio is of special interest in such situations. Our results show that the proposed approach is robust against low transcription performance on one side of the automatically transcribed pSp corpus, confirming the approach's feasibility in the context of resource limitation. We consider the result that pSp audio can replace as well as enhance parallel text data as training resource for translation model development as our most important finding, 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 pSp speech potentially includes high amounts of out-of-vocabulary words. The magnitude of the OOV problem depends strongly on the domain at hand and the vocabulary size of the used ASR systems, as for example shown by Hetherington (1995). Optimizing the vocabulary size to the current domain is one solution, but this may be hard to accomplish in the context of ASR of 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. (2006) propose an interesting solution to this problem. They propose to apply phone based ST where TMs are learned on a parallel corpus of foreign phone sequences and corresponding English translation.

We believe that the question of how to mix traditional training data with pSp for a maximal gain in performance, at an acceptable cost level, deserves more attention. For example, one could ask if it makes more sense to transcribe higher amounts of speech data for reduced WER 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 WER or content are of special interest.

We further believe that future work has to address larger amounts of pSp audio and more language pairs, to further support our hypotheses regarding the translation performance of pSp-trained TMs. While the attached collection effort of additional pSp audio can 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 pSp 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 CI 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.

Acknowledgements

This work was partly funded by the European Union under the project TC-STAR (IST-2002-FP6-506738), the National Science Foundation (NSF) under the project STR-DUST (IIS-0325905) and by the US DARPA under the project TRANSTAC.

References

Al-Kahnjii, R., El-shiyab, S., Hussein, R., 2000. On the use of compensatory strategies in simultaneous interpretation. Meta: Journal des traducteurs 45 (3), 544–557.

Besacier, L., Zhou, B., Gao, Y., 2006. Towards Speech Translation of Non Written Languages. SLT, Palm Beach, Aruba.

Besien, F.V., 1999. Anticipation in simultaneous interpretation. Meta: Journal des traducteurs 44 (2), 250–259.

Hetherington, I.L., A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Kemp, T., Waibel, A., 1999. Unsupervised Training of a Speech Recognizer: Recent Experiments. Eurospeech, Budapest, Hungary.

Khadivi, S., Zolnay, A., Ney, H., 2005. Automatic Text Dictation in Computer-assisted Translation. Interspeech, Portugal, Lisbon.

Koehn, P., Monz, C., 2006. Manual and automatic evaluation of machine translation between European languages. In: Proc. on the Workshop on Statistical Machine Translation , New York City, USA, pp. 102–121.

Kohn, K., Kalina, S., 1996. The strategic dimension of interpreting. Meta: Journal des traducteurs 41 (1), 118–138.

Lamel, G.A.L., Gauvain, J., 2002. Unsupervised Acoustic Model Training. ICASSP, Orlando, USA.

Lamel, L., Gauvain, J., Adda, G., 2001. Investigating Lightly Supervised Acoustic Model Training. ICASSP, Salt Lake City, USA.

Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Computer Speech and Language, 373–400.

Matusov, E., Leusch, G., Bender, O., Ney, H., 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. IWSLT, Pittsburgh, USA, pp. 148–154.

Mostefa, D., Hamon, O., Moreau, N., Choukri, K., May 2007. TC-STAR Evaluation Report, Del.no 30, http://www.tc-star.org.

Och, F., 2003. Minimum error rate training in statistical machine translation. In: Proc of ACL , Sapporo, Japan.

Och, F., Ney, H., 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29 (1), 19–51.

Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. ACL, Philadelphia, PA, USA.

Paulik, M., Waibel, A., 2008a. Lightly Supervised Acoustic Model Training on EPPS Recordings. Interspeech, Brisbane, Australia.

Paulik, M., Waibel, A., 2008b. Extracting clues from human interpreter speech for spoken language translation. In: Proc. of ICASSP , Las Vegas, NV, USA.

Paulik, M., Waibel, A., 2009. Automatic Translation from Parallel Speech: Simultaneous Interpretation as MT Training Data. ASRU, Merano, Italy.

Paulik, M., Waibel, A., 2010. Spoken Language Translation from Parallel Speech Audio: Simultaneous Interpretation as SLT Training Data. ICASSP, Dallas, TX, USA.

Reddy, A., Rose, R., 2008. Towards Domain Independence in Machine Aided Human Translation. Interspeech, Brisbane, Australia.

Santiago, R., http://home.earthlink.net/ terperto/id16.html (accessed 11.01.10.) 2004. Consecutive Interpreting: A Brief Review.

Seleskovitch, D., 1978. Interpreting for International Conferences. Pen and Booth, Arlington, VA.

Smith, F., 1985. Reading Without Nonsense. NY TeacherÕs College Press, NY.

Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Intl. Conf. on Spoken Language Processing , Denver, CO, USA.

Vogel, S., 2003. SMT Decoder Dissected: Word Reordering. In: Proc. of Coling , Beijing, China.

Matthias Paulik received his Masters degree (Dipl.-Inform.) and Ph.D. degree (Dr.-Ing., *summa cum laude*) in Computer Science from Karlsruhe Institute of Technology (KIT), Germany in 2005 and 2010, respectively. During his doctoral studies (2005–2010) he worked as a staff member at Carnegie Mellon University (CMU) and part-time as a research scientist at Mobile Technologies, LLC. Since summer 2010, he works at Cisco's Speech and Language Technology (C-SALT) team where he conducts product driven research and development in the areas of speech and language technology. His research interests are centered around automatic speech recognition, statistical machine translation and the combination of these two technologies for the purpose of automatic speech(-to-speech) translation.

Alex Waibel is a professor of computer science at CMU, USA and at the KIT, Germany. He directs the international Center for Advanced Communication Technologies (interACT) at both locations with research interests in multimodal and multilingual human communication systems. His team pioneered many of the first domain-limited and domain-unlimited speech translators. He was one of the founders and chairmen (1998–2000) of C-STAR, the consortium for speech translation research. He published extensively in the field, received several patents and awards, and launched several successful companies. He received his B.S., M.S., and Ph.D. degrees at Massachusetts Institute of Technology and CMU, respectively.