# On the Integration of Speech Recognition and Statistical Machine Translation

*E. Matusov, S. Kanthak, and H. Ney*

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University
52056 Aachen, Germany
{matusov,kanthak,ney}@informatik.rwth-aachen.de

## Abstract

This paper focuses on the interface between speech recognition and machine translation in a speech translation system. Based on a thorough theoretical framework, we exploit word lattices of automatic speech recognition hypotheses as input to our translation system which is based on weighted finite-state transducers. We show that acoustic recognition scores of the recognized words in the lattices positively and significantly affect the translation quality. In experiments, we have found consistent improvements on three different corpora compared with translations of single best recognized results. In addition we build and evaluate a fully integrated speech translation model.

## 1. Introduction

It has been shown in the past that automatic speech recognition (ASR) and machine translation (MT) can be coupled in order to directly translate spoken utterances into another language. However, we find that previously presented work does not cover the topic completely and partially comes to contradictory or wrong conclusions.

Various approaches to speech translation have been proposed and investigated by [1], [2], [3] and more recently [4]. [1] presents an integrated speech translation system for tasks from the Eutrans project. However, the experimental results were inconsistent as integrated speech translation performed better than the serial approach on data with an artificially generated bilingual corpus, while it performed much worse on real-world data. [2] presents only the theory of integrated speech translation, but lacks experimental results. More recently, [4] concludes that improvements from tighter coupling may only be observed when lattices are sparse, i.e. when there are only few hypothesized words per spoken word in the lattice. This is inconsistent with the theory of [2] and would mean that integrated speech translation would not work at all.

Following [2], Section 2 reviews the Bayes' decision rule for speech translation. Starting from there we show how to integrate the translation and the acoustic model. For the translation model we propose an alternative view on word alignment and monotonization. Furthermore, we improve the training procedure by taking advantage of the word alignment information for reordering of the target sentences in the training corpus. The translation model is then implemented efficiently using a generic finite-state toolkit which supports on-demand computation [5]. In order to improve the translation quality, we directly translate from ASR word lattices and consistently benefit from acoustic scores. We thoroughly analyze the dependence of word lattice density on the translation quality and show that lattices with higher densities improve translation error measures. This leads to the conclusion that fully integrated speech translation should work and we prove this on the Eutrans II task.

## 2. Bayes' Decision Rule for Speech Translation

In speech translation, we are looking for a target language sentence $e_1^I$ which is the translation of a speech utterance represented by acoustic vectors $x_1^T$. In order to minimize the number of sentence errors, we maximize the posterior probability of the target language translation given the speech signal (see [2]):

$$\begin{aligned}
\hat{e}_1^{\hat{I}} &= \operatorname*{argmax}_{I,e_1^I} Pr(e_1^I|x_1^T) \\
&= \operatorname*{argmax}_{I,e_1^I} Pr(e_1^I) \cdot Pr(x_1^T|e_1^I) \\
&= \operatorname*{argmax}_{I,e_1^I} Pr(e_1^I) \cdot \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J, e_1^I) \\
&\cong \operatorname*{argmax}_{I,e_1^I} Pr(e_1^I) \cdot \max_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J) \\
&= \operatorname*{argmax}_{I,e_1^I} \max_{f_1^J} Pr(e_1^I) \cdot Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J) \\
&= \operatorname*{argmax}_{I,e_1^I} \max_{f_1^J} Pr(f_1^J, e_1^I) \cdot Pr(x_1^T|f_1^J)
\end{aligned}$$

Note that we made the natural assumption that the speech signal does not depend on the target sentence and approximated the sum over all possible source language transcriptions by the maximum. $Pr(f_1^J, e_1^I)$ refers to the translation model while $Pr(x_1^T|f_1^J)$ may be a standard acoustic model. The translation model is used instead of the usual language model.

Following [6], we introduce an *alignment* between a source and target sentence as the hidden variable $\mathcal{A}$:

$$Pr(f_1^J, e_1^I) = \sum_{\mathcal{A}} Pr(\mathcal{A}) \cdot Pr(f_1^J, e_1^I|\mathcal{A})$$

The hidden alignment $\mathcal{A}$ represents all possible interpretations of source words by target words.

In order to estimate a finite-state transducer model for translation, we aim at a representation that, by following a path from the initial state of the transducer to one of its final states labeled with the source words $f_1^J$, translates into a sequence of target words $e_1^I$ of length $I$ possibly different from $J$. Therefore, we restrict the hidden alignment $\mathcal{A}$ by allowing only alignments in which each target word is only connected to one source word. The alignment can be represented as a function $a : \{1, ..., I\} \rightarrow \{1, ..., J\}$. Due to this alignment definition, some positions $j$ may remain unaligned. We then additionally modify the alignment function to be *monotonic*, i. e. for each pair of target positions $i < i'$ we require that $a_i \leq a_{i'}$.

We use the GIZA++ toolkit [7] to automatically train word alignment functions. For two languages with similar word order

these alignments are usually monotonic. In general, we apply the GIATI monotonization technique described in [8] to force a monotonic alignment. Using the monotonic alignment function $a$ we then define the segmentation of $e_1^I$ into target phrases $\tilde{e}_j$ as follows ($l \geq 0$):

$$\tilde{e}_j = \{e_i, e_{i+1}, ..., e_{i+l} | a_i = a_{i+1} = ... = a_{i+l} = j\}$$

The word sequence $\tilde{e}_j$ is unique due to the alignment monotonicity. For source positions $j$ with no alignment we set $\tilde{e}_j = \varepsilon$.

If we additionally assume that probabilities only depend on the immediate predecessor words, we can compute the *translation model* $Pr(f_1^J, e_1^I|\mathcal{A})$ as:

$$Pr(f_1^J, e_1^I|\mathcal{A}) \cong \prod_{j=1}^{J} p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}, a) \qquad (1)$$

Thus, the translation model is a statistical $m$-gram language model on the level of bilingual (source word, target phrase)-pairs $(f_j, \tilde{e}_j)$ and well-known smoothing techniques may be used for better generalization. An example of a transformed corpus of bilingual tuples is given in Figure 1.

At present, we do not explicitly model the alignment probability $Pr(\mathcal{A})$. Instead we search for the alignment (i. e. for a segmentation into sequences $\tilde{e}_j$) which maximizes the translation model probability as given by Equation 1. Finally, we arrive at the following search criterion:

$$\hat{e}_1^{\hat{I}} = \underset{\tilde{e}_1^J, a}{\operatorname{argmax}} \max_{f_1^J} \prod_{j=1}^{J} p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}, a) \cdot p(x_{t_{j-1}}^{t_j} | f_j)$$

It should be emphasized here that the joint $m$-gram probability $p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}, a)$ contains dependencies on the predecessor source words $f_{j-m}^{j-1}$ and therefore also serves as a source language model.

The optimization criterion above suggests that we use word lattices as input for speech translation. Each arc in the lattices is scored with the conditional probability of the acoustic signal given a source word hypothesis:

$$Pr(x_1^T | f_1^J) \quad \cong \quad \prod_{j=1}^{J} p(x_{t_{j-1}}^{t_j} | f_j)$$

where $t_{j-1}$ is the starting time and $t_j$ is the end time of the hypothesized acoustic realization of $f_j$.

### 2.1. Reordering in Training and Translation

In order to eliminate the monotonization heuristics used in the GIATI approach, we reorder the sentences in the target training corpus based on the alignment $a$ such that the resulting alignment becomes monotonic. Obviously, resulting translations will have the word order of the source sentence. For languages with similar order, this is not necessarily harmful. To fix the wrong word order in general, we use a similar idea to that described in [3]. Given a single best translation result we first permute the words which results in a permutation automaton. The number of possible permutations can be reduced to a reasonable amount using IBM constraints together with a fixed window size. In addition, we define a probability distribution over the permutation automaton that favors the original order. We build the resulting permutation automaton on-demand and compose it with an $n$-gram target language model in order to select the word order with the highest probability. Unless otherwise specified, we will follow this novel approach in our experiments.

`vorrei|i'd_like del|some gelato|ice_cream per|`$\varepsilon$ `favore|please`

Figure 1: Example of a transformed sentence pair.

## 3. Experimental Results

### 3.1. Corpus Statistics

The speech translation experiments were carried out on three different tasks. Experiments for all tasks were based on bilingual sentence-aligned corpora. Corpus statistics for these tasks are given in Table 1.

The Italian-English *Basic Travel Expression Corpus* (BTEC) task contains tourism-related sentences usually found in phrase books for tourists going abroad. We were kindly provided with this corpus by ITC-IRST. 16 reference translations of the correct transcriptions of this corpus were available.

The Italian-English Eutrans II FUB task contains sentences from the domain of hotel help-desk requests. It is significantly smaller than the BTEC task and has evolved from one of the first European-funded speech translation projects.

The third task is a Spanish-English and Spanish-Catalan translation task. The available bilingual corpora were prepared within the European LC-STAR project [9] and contain spontaneous utterances from travel and appointment scheduling domains. These utterances are complete telephone dialogue turns and significantly longer (about 23 words on average) than the utterances in the BTEC corpus.

### 3.2. Model Scaling Factors

Since both the acoustic model probabilities $p(x_{t_j}^{t'_j}|f_j)$ and the joint translation probabilities $p(f_j, \tilde{e}_j | f_{j-m}^{j-1}, \tilde{e}_{j-m}^{j-1}, \mathbf{a})$ are only approximations of the true distributions, we add a scaling exponent $\lambda$ to the translation model. Except for the Eutrans II task, we optimized the scaling factor $\lambda$ on a development set, which was similar in size to the corresponding test set.

### 3.3. Evaluation Criteria

For the automatic evaluation, we used word error rate (WER), position-independent word error rate (PER), the BLEU and NIST scores [10, 11]. The latter two measure accuracy, i. e. larger scores are better. The error rates and scores were computed with respect to multiple reference translations when available. To indicate this, we will label the error rate acronyms with an $m$. On the BTEC Italian-English task, both training and evaluation were performed using the corpus and references in lowercase. On all tasks, both training and evaluation were carried out without punctuation marks.

### 3.4. Generation of Word Lattices

The speech recognition systems used here produce word lattices where arcs are labeled with start and end time, the recognized entity (word, noise, hesitation, silence), the negative log probability of acoustic vectors between start and end time given the entity and the negative log language model probability of the entity. In a first step we mapped all entities that were not spoken words onto the empty arc label $\varepsilon$. As language model probabilities and the time information are not used in our approach, we removed them from the lattices and compressed the structure by applying $\varepsilon$-removal and determinization. As most of the translation experiments were done without pruning, this step significantly reduced runtime without changing the results.

Table 1: Corpus statistics of the speech translation tasks BTEC, Eutrans II and LC-STAR.

| | | BTEC | | Eutrans II FUB | | LC-STAR | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Italian | English | Italian | English | Spanish | English | Spanish | Catalan |
| Train | Sentences | 66107 | | 3257 | | 39018 | | 41885 | |
| | Running Words | 410275 | 427402 | 47681 | 57663 | 427014 | 456198 | 534215 | 544731 |
| | Vocabulary | 15983 | 10971 | 2453 | 1695 | 10821 | 9303 | 11834 | 12163 |
| | Singletons | 6386 | 3974 | 975 | 519 | 3661 | 3660 | 4216 | 4191 |
| Test | Sentences | 253 | | 300 | | 519 | | 519 | |
| | Running Words | 1459 | 1510 | 5305 | 6419 | 13365 | 14200 | 13365 | 13485 |
| Out-Of-Vocabulary rate [%] | | 2.5 | 0.9 | 2.3 | 1.3 | 2.2 | 2.2 | 1.3 | 1.3 |
| ASR WER [%] | | 21.4 | - | 23.7 | - | 31.9 | - | 31.9 | - |

Table 2: Translation results for the BTEC Italian-English task.

| Input/ transcription: | mWER [%] | mPER [%] | BLEU [%] | NIST |
|---|---|---|---|---|
| correct text | 25.7 | 20.2 | 61.3 | 9.94 |
| single best | 33.2 | 28.4 | 53.1 | 8.91 |
| word lattice | 31.9 | 27.4 | 54.9 | 8.96 |
| + ac. scores | 30.6 | 26.0 | 55.4 | 9.17 |

Table 3: Effect of target reordering in training and after translation for word lattice translation on the BTEC task.

| Type of reordering: | mWER [%] | mPER [%] | BLEU [%] | NIST |
|---|---|---|---|---|
| none | 31.6 | 27.6 | 54.3 | 8.95 |
| target | 30.6 | 26.0 | 55.4 | 9.17 |

### 3.5. BTEC Italian-English Task

On the BTEC task, the best translation results were obtained by estimating a smoothed $4$-gram language model on the level of bilingual tuples $(f_j, \tilde{e}_j)$. When translating, we performed full search, even when using word lattices as input. We also used a $4$-gram target language model to score and select constrained reorderings of the produced translations.

The experimental results for the BTEC test corpus are given in Table 2. When translating single best recognition results instead of correctly transcribed ones, the quality of machine translation degrades by about 23 % relative in WER.

In the next experiment, we compose the word lattice containing multiple hypotheses of the recognized utterances with the translation transducer. First, we do not use acoustic scores of the labels in the input lattice, i. e. we only exploit the lattice topology. We see that the error measures slightly improve.

Next, we use the acoustic scores of the word lattice together with translation model scores in the global decision process. On the BTEC task, the optimal translation model scaling factor $\lambda$ was found to be $45$. With this setting, the translation quality was significantly improved, both on the development corpus and the test corpus (in Table 2 the mWER drops from 33.2 to 30.6 %).

In a contrasting experiment, we kept the original word order in the target training sentences and re-estimated the $4$-gram translation model (using the GIATI technique). Table 3 shows that the model without reordering performs significantly worse. Also, without monotonization the optimal scaling factor for the translation model scores was found to be even higher, $\lambda = 55$.

A possible explanation for the high translation model scaling factors is the fact that, in contrast to speech recognition, the decision concerning the target sentence does not rely as much on the acoustics as it does for the source sentence. When we apply monotonization, the target language structure (e. g. word order) gets closer to the structure of the source language, which in turn results in a smaller optimal scaling factor and better translation quality.

### 3.6. FUB Italian-English Task

In contrast to the experiments on the BTEC task, only a limited amount of training data was available for the FUB task. Since a bigram translation model yielded the best WER for written text input we used a bigram for all experiments on the FUB task.

We had more control over the recognition experiments here and generated lattices with different densities. The lattice error rate, i.e. the minimum WER among all paths through the lattice, was as low as 9.1% for the largest lattice density of 2098. We optimized the system with respect to both the lattice density and the translation model scaling factor $\lambda$ simultaneously. Figure 2 shows the effect on the translation WER. In contrast to the results of [4], the WER consistently drops with larger lattices and shows a clear minimum for $\lambda = 90$. Results of all error measures for the optimal settings are given in Table 4.

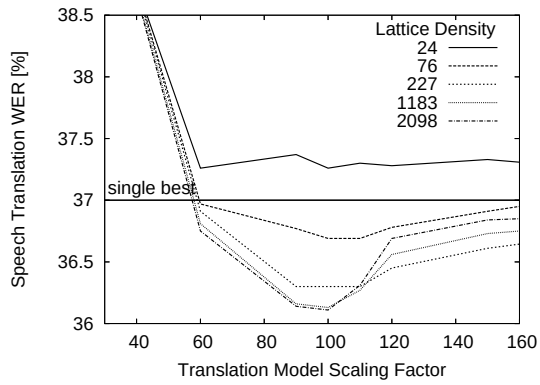


Figure 2: Effect of different lattice densities and translation model scaling factors $\lambda$ on the translation WER (FUB task). Lattices with density 24 were generated using tight beams which resulted in a slightly worse recognition WER.

Additionally, we coupled speech recognition and machine translation directly using a single finite-state network. Building the network follows the description of [12] where we substituted the language model by the translation model. The last line of Table 4 shows that the fully integrated system performs only slightly worse than the system using large lattices which we explain by a couple of search errors and the fact that we did not optimize the search network on state-level due to technical problems. Note, that although the speech recognition system has a slightly worse WER on this task compared to [13] we obtain a much better speech translation WER. The integrated system also gets better compared to using only the single best recognition output. While the improvements for the different error measures are quite small on this task, we also calculated the WER probability of improvement (POI) [14]. The POI for the lattice-based approach was 96.2% which emphasizes that even these small improvements are significant.

Table 4: Translation results on the FUB Italian-English task. The last line contains results when directly coupling the speech recognition and machine translation systems by using a single optimized finite-state network.

| Input/ transcription: | WER [%] | PER [%] | BLEU [%] | NIST |
|---|---|---|---|---|
| correct text | 28.8 | 21.8 | 59.2 | 8.35 |
| single best | 37.0 | 28.7 | 51.8 | 7.36 |
| word lattice | 38.8 | 30.6 | 48.9 | 7.23 |
| + ac. scores | 36.1 | 28.1 | 52.7 | 7.45 |
| integrated | 36.4 | 29.2 | 52.2 | 7.40 |

### 3.7. LC-STAR Spanish-English and Spanish-Catalan Tasks

On the LC-STAR task, the Spanish speech recognition system had an accuracy of less than 70 % (see Table 1) due to limited training data, low sampling rate and large speaker variability. When the highly erroneous single best recognition hypotheses are translated into English, the WER climbs over 60 %, see Table 5. The translation quality can be improved only slightly when we translate high-density word lattices with acoustic scores. We attribute this to the undertrained translation model (the error rates are already rather high for the correct Spanish input). This translation model can not discriminate well between correct and erroneous hypotheses in the word lattice. This is supported by the following observation. When we use exactly the same Spanish word lattices and translate to Catalan, we reach an improvement of over 12 % relative in WER over the translation of the single best input (Table 5). Here, the translation model is more robust since the difference in structure and word order between Spanish and Catalan is much smaller than between Spanish and English. The optimal scaling factor $\lambda$ for the translation model scores was determined to be 60 on a development set in both of these experiments.

Table 5: Translation results for the LC-STAR Spanish-English and Spanish-Catalan tasks. Lattices contain acoustic scores.

| | Input: | WER [%] | PER [%] | BLEU [%] | NIST |
|---|---|---|---|---|---|
| Spanish to English | correct text | 44.2 | 32.5 | 37.2 | 8.00 |
| | single best | 60.6 | 45.5 | 25.9 | 6.22 |
| | word lattice | 59.8 | 45.2 | 25.6 | 6.30 |
| Spanish to Catalan | correct text | 12.2 | 10.5 | 80.1 | 12.03 |
| | single best | 39.8 | 32.3 | 47.6 | 8.56 |
| | word lattice | 34.9 | 28.7 | 53.7 | 9.21 |

## 4. Conclusions

In this paper, we have used ASR word lattices as input for a statistical translation system. Coupling of speech recognition and machine translation was implemented efficiently with weighted finite-state transducers. By using word lattices with acoustic model scores instead of single best recognition results, we were able to avoid the negative impact of recognition errors and consistently improved translation quality on three different tasks. We also proposed and implemented word reordering for target sentences both in training and after translation, and further improved the translation results. In contrast to previously published work, for the first time we were able to gain improvements even with large lattices and a non-serial, integrated speech translation approach. In the future, we plan to more extensively explore lexical reordering and test our system on larger tasks.

## 5. Acknowledgement

This work was in part funded by the European Union under the project LC-STAR, IST-2001-32216, and under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

## 6. References

[1] E. Vidal, "Finite-State Speech-to-Speech Translation", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 111–114, Munich, Germany, 1997.

[2] H. Ney, "Speech Translation: Coupling of Recognition and Translation", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1149–1152, Phoenix, AZ, 1999.

[3] S. Bangalore and G. Riccardi, "Finite-State Models for Lexical Reordering in Spoken Language Translation", Proc. Int. Conf. on Spoken Language Processing, vol. 4, pp. 422–425, Beijing, China, 2000.

[4] S. Saleem, S.-C. Jou, S. Vogel, and T. Schultz, "Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems", Proc. Int. Conf. on Spoken Language Processing, pp. 41–44, Jeju Island, Korea, 2004.

[5] S. Kanthak and H. Ney, "FSA: An Efficient and Flexible C++ Toolkit for Finite State Automata using On-demand Computation", Proc. 42nd Annual Meeting of the ACL, pp. 510 – 517, Barcelona, Spain, 2004.

[6] P. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation", Computational Linguistics, vol. 19(2):263–311, 1993.

[7] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, vol. 29, number 1, pp. 19–51, 2003.

[8] F. Casacuberta and E. Vidal, "Machine Translation with Inferred Stochastic Finite-State Transducers", Computational Linguistics, vol. 30(2):205-225, 2004.

[9] V. Arranz, N. Castell, J. Giménez, "Development of Language Resources for Speech-to-Speech Translation", Poster, RANLP2003, Borovets, Bulgaria. http://www.lc-star.com, 2003.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proc. 40th Annual Meeting of the ACL, Philadelphia, PA, pp. 311–318, 2002.

[11] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using n-gram Co-Occurrence Statistics", Proc. Human Language Technology Conf., San Diego, CA, 2002.

[12] M. Mohri, F. C. N. Pereira and M. Riley, "Weighted Finite-State Transducers in Speech Recognition", Proc. ISCA Workshop, ASR2000, Paris, France, 2000.

[13] F. Casacuberta, D. Llorens, C. Martínez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Picó, A. Sanchis, E. Vidal, and J. M. Vilar, "Speech-to-speech Translation Based on Finite-State Transducers", Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 613–616, Salt Lake City, UH, 2001.

[14] M. Bisani, and H. Ney, "Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation", Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Montreal, Canada, Vol. 1, pp. 409–412, 2004.