Translating language with technology's help

MATTHIAS PAULIK, SEBASTIAN STÜKER, CHRISTIAN FÜGEN, TANJA SCHULTZ, AND ALEX WAIBEL

IN HUMAN-MEDIATED TRANSLATION scenarios, a human interpreter translates from either a spoken or a written representation of a source language into a target language. In the European Parliament, for example, interpreters simultaneously translate the speech of a Spanish speaker into the languages of the listeners.

Sometimes, human interpreters can also use additional textual information, e.g., the manuscript of a speech, to help improve their translation. In many scenarios, it is desirable to have a transcript of the original speech along with the simultaneous translations for archiving or publication purposes. Here, automatic transcription systems can help in lowering costs and the effort needed to obtain such transcripts. Our work is aimed at improving these automatic transcription systems over the current state-of-the-art systems, which are still error prone.

Automatic speech recognition, machine and speech translation

Automatic speech recognition (ASR) is a pattern recognition problem with the purpose of finding the word sequence W that belongs to a given audio recording of speech, the pattern X. Due to the variability that is inherent in human speech, this classification process is often done with the help of statistical models. Stating the ASR problem in terms of probability theory and applying Bayes' theorem to it leads to the fundamental equation of speech recognition

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W|X) \\ &= \arg\max_{W} \frac{P(W)P(X|W)}{P(X)} \\ &= \arg\max_{W} P(W)P(X|W). \end{aligned} \quad (1)$$

In other words, speech recognition finds the most probable word sequence $\hat{W}$ given the observed pattern X extracted from the recorded speech. To do so, the product of $P(W)$ and P(X|W) has to be maximized. This process is usually called decoding or search, since it can be interpreted as searching for the best word sequence in the hypotheses space of all possible sequences.

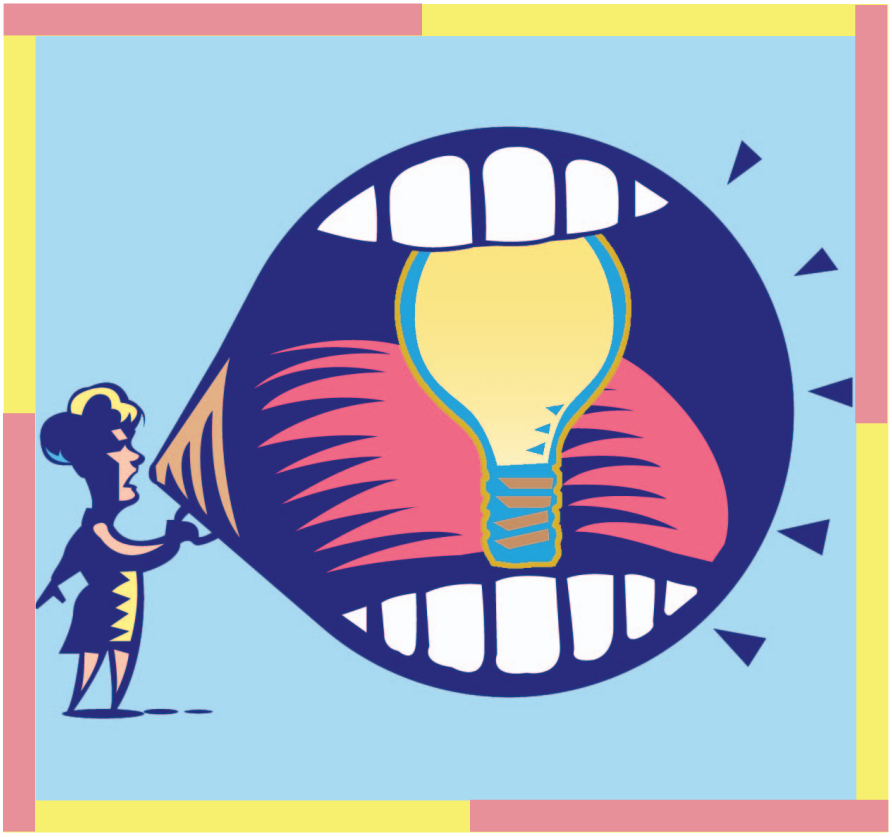
$P(W)$ is called the *language model* (LM) and determines the prior probability that the word sequence W is spoken. P(X|W) is referred to as the *acoustic model* (AM) and links the spoken words to the acoustic manifestation of speech.

Machine translation (MT) is the problem of automatically translating text from a source language, let's say Spanish, to a target language, e.g., English. Current state-of-the-art systems utilize statistical models to solve this problem as well. Similar to the fundamental equation of speech recognition, one can formulate the MT problem as

$$\begin{aligned}\hat{T} &= \arg\max_{T} P(T|S) \\ &= \arg\max_{T} \frac{P(T)P(S|T)}{P(S)} \\ &= \arg\max_{T} P(T)P(S|T). \end{aligned} \quad (2)$$

$P(T)$ again is called the (target) LM, while P(S|T) is called the *translation model* (TM).

The AM is trained with the help of manually transcribed speech data. Similarly, the TM is trained on sample translations produced by human interpreters. The LMs are estimated on large amounts of monolingual text data. The most widely used type of LM is the so



0278-6648/07/$25.00 © 2007 IEEE

called *n*-gram LM, with *n* usually in the order of three (tri-gram LM) or four. Here, the probability of a word depends only on the history of the *n*–1 preceding words. However, today's ASR and MT systems are not perfect but rather produce error prone output of varying quality.

When we combine ASR with MT, we get *speech translation systems* (STS) that translate speech into a different language.

Machine translation enhanced automatic speech recognition

A straightforward approach to automate the transcription process in human-mediated speech translation scenarios is to simply apply ASR to the speech of the interpreters and speakers. For the transcription of the translator's speech, however, additional knowledge in the form of the source language that is being translated is available and can be used to improve the quality of the speech recognition system. One way to achieve this is to use MT to translate these resources from the source into the target language. The ASR system can then be biased towards the knowledge gained. We call the process of recognizing speech using a system that has been improved in this way *MT enhanced ASR* (MTEASR).

Previous work in MTEASR considered the case of a written source language representation as proposed in 1994 by Dymetman et al. and by Brown et al. In the TransTalk project, Dymetman and his colleagues improved the automatic transcription of a translator's speech by rescoring the ASR *n*-best lists (the *n*: most likely hypotheses for a given utterance) with a TM. Further, they used the TM to dynamically create a sentence-based vocabulary list to restrict the ASR search space. Brown et al. introduced a technique for applying the translation model during decoding by combining its probabilities with those of the LM.

Our work goes beyond the described research by proposing an iterative system that incorporates all knowledge sources available for both—the source and target language—into an integrated system. Figure 1 depicts the overall design in the case of a spoken source language representation. In this scenario, a Spanish talker's speech is being translated by a human interpreter into English. At the same time, a Spanish ASR system is transcribing the Spanish speech, and an MT system is automatically translating the result into English—parallel to the human interpreter. The automatic translation is used to improve the automatic transcription of the interpreter. This transcript is then being translated back into Spanish by an automatic system. This translation is then in turn used to adapt the recognition system toward the Spanish speaker. Both automatic transcriptions in parallel can further be used to improve the MT systems, completing one cycle of our iterative design. The now-improved transcription of the Spanish speaker together with the improved MT component can be used to enter a new iteration of improving all components.

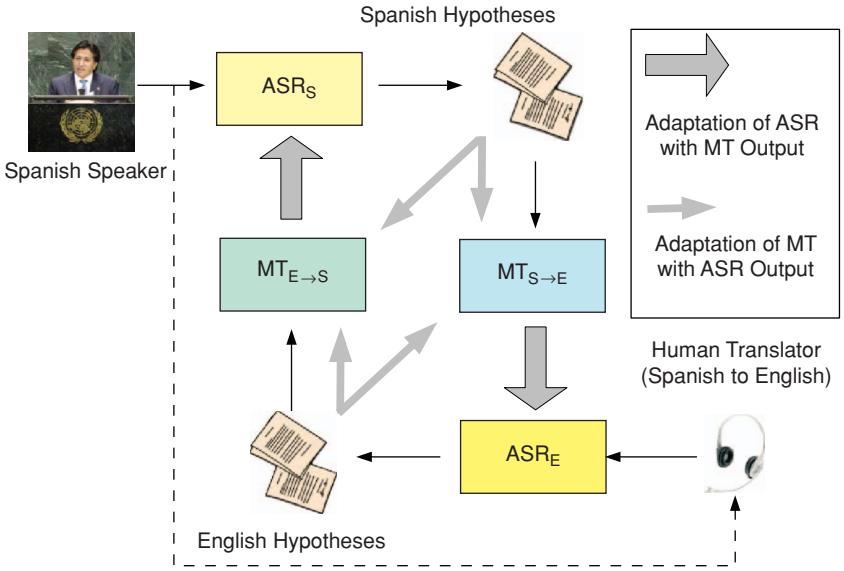


Fig. 1 MTEASR in case of a spoken source language representation

Task and baseline components

Task and data sets

We performed our experiments in the domain of basic tourist phrases as they can be found, for example, in the basic travel expression corpus (BTEC). BTEC contains tourist phrases and expressions that cover the basic needs of a traveler as one would find them in commercial phrase books. For development and evaluation, we used two different data sets. On Data Set I, which consists of 506 parallel Spanish and English sentences, we evaluated several basic adaptation techniques in the case of a written source language representation. The sentences were each read four times by a total of 12 different speakers. After removing corrupted recordings, 2,008 spoken utterances amounting to 67 min of speech remained. On Data Set II, we evaluated our iterative system design in the case of a written as well as in the case of a spoken source language representation. Data Set II consists of 500 parallel English and Spanish sentences in form and content close to BTEC. The sentences were read twice. Ten percent of the data was randomly selected as held-out data for system parameter tuning. Due to some flawed recordings, the English data set contains 880 sentences, while the Spanish data set consists of 900 sentences. The Spanish audio data equals 45 min while the English equals 33 min of speech.

Speech recognition system

For the ASR experiments in this work, we used the Janus recognition toolkit (JRTk), featuring the Ibis single pass decoder, named after the Egyptian god of writing, reckoning, and learning. Ibis is a time-synchronous decoder that, unlike our earlier decoder, allows us to apply full LM information at a very early stage in the decoding process in a single pass over the audio to be recognized. IBIS generates word graphs as a result of its search, which can be rescored using a different LM than during decoding and can be used to derive the *n*-most likely recognition hypotheses. The AMs of the English recognizer were

trained on 180 h broadcast news data and 96 h meeting data. The Spanish system was trained on 112 h South American speech data (mainly Mexican and Costa Rican dialects) and 14 h Castilian speech data. The back-off trigram language models of the two speech recognizers were trained on the respective English/Spanish part of the bilingual BTEC. The term "back off" refers to the fact that, should the trigram LM not directly contain the probability of a word given its two-word history, the probability is estimated on the shorter history of the one or zero preceding words. Table 1 gives an overview of the word error rate (WER) on the two data sets along with the out-of-vocabulary (OOV) rates and LM perplexities. The LM perplexity (PPL) can be interpreted as the average number of words from which the next word is being chosen given the preceding words. A lower PPL indicates an easier choice for the recognizer. Thus, all other things being equal, using an LM with a lower PPL normally leads to a lower WER. The comparison of the WER between different languages is complicated due to major differences in the complexity on all linguistic levels.

Table 1. Word error rate, out-of-vocabulary rate, and LM perplexity (PPL) of the baseline ASR systems

	Data Set I —English	Data Set II —English	Data Set II —Spanish
WER (%)	12.6	20.4	17.2
OOV (%)	0.52	0.53	2.04
PPL	21.6	86.0	130.2

However, the much higher English WER on Data Set II compared to Data Set I can be explained by the LM PPL on Data Set II, which is approximately four times higher.

Machine translation system

The Interactive systems Labs (ISL) statistical MT system was used for the English to Spanish and the Spanish to English automatic translations. The translation systems for both directions were trained on the bilingual Spanish/English BTEC. The ISL statistical MT system produces an n-best list of translation hypotheses for a given source sentence with the help of its TM, target LM, and translation memory. The translation memory searches for each source sentence that has to be translated for the closest matching source sentence with regard to the edit distance in the training corpus and extracts it along with its translation. The performance values of the baseline MT systems on the two data sets and the two conditions (written and spoken source language representation) are listed in the following sections along with the (iterative) MTEASR experiment results.

Basic adaptation techniques

In this section, we compare different basic adaptation techniques for improving the performance of the system's main components on the basis of a written source language representation. In particular, we describe techniques to adapt the ASR component using knowledge provided by the MT component and techniques to adapt the MT component using knowledge derived from ASR. The performance improvements on the ASR are described in terms of WER and were obtained by using the baseline MT knowledge only. For the experiments on the MT component the improved output of the adapted ASR was used. The MT performance improvements are reported in the bilingual evaluation understudy (BLEU) in respect to one reference translation. The BLEU score ranges from 0–100, whereas a translation that is identical to its reference translation attains a score of 100. The score is computed as the geometrical mean of the n-gram precisions with $n \in \{1;2;3;4\}$, i.e., it measures the n-gram cooccurrence between a given translation and one or more reference translations. All experiments in this section were conducted on Data Set I.

ASR adaptation techniques

Hypothesis selection by rescoring

The n-best WER (nWER) found within the English ASR 150-best lists (the lists of the 150 best hypotheses for each sentence to be recognized) of the baseline system is 6.5% compared to 12.6% for the first best result, thus indicating huge potential for rescoring the ASR n-best lists. In contrast, the best WER that can be achieved on the 150-best MT list is 34.2%. However, when combining the n-best lists of ASR and MT, the nWER drops to 4.2%, which proves that complementary information is given in the n-best lists of both components. In fact, we observed the best rescoring performance when enriching the ASR 150-best list with just the first best MT hypothesis. Therefore, our reported rescoring results refer to ASR n-best lists enriched in this manner. The rescoring algorithm that we applied computes new scores (negative log probabilities) for each sentence by adding up the weighted and normalized TM score, LM score, and ASR score of this sentence. In addition, the rescoring algorithm uses word context classes: MT monograms, trigrams, and complete MT sentences. MT n-grams are n-grams included in the respective MT n-best list; MT sentences are defined in the same manner. Whenever one of these word context classes is found within an ASR hypothesis by the rescoring algorithm, the score of the hypothesis is improved by a value specific to the respective word context class. Parameter optimization was done by manual gradient descent. The resulting best system yielded a WER of 10.5%, which corresponds to a relative error rate reduction of 16.7%. We found that the MT monogram discounts have the strongest influence on the success of this approach, followed by the TM score. This suggests that the MT is not very useful in getting additional word context information in the form of MT bigrams and trigrams but very useful as a provider for a bag of words that predicts which words are going to be said by the human translator. This approach offers a successful way to apply MT knowledge for ASR improvement without changing the ASR system.

Cache language model

Since the monogram discounts have such a great impact on the success of the rescoring approach, it is desirable to use this form of MT knowledge not only after but during ASR decoding. In our cache LM approach, we define the members of the word-class monogram in the same manner as above, but instead of rescoring n-best lists, we now modify the score of the ASR hypotheses during decoding. The best performing system yielded a WER of 10.4%, and

therefore had a similar performance as the rescoring approach although it lacks the direct computation of the TM score. This can be explained by the fact that the expectation to find new, correct hypotheses could be fulfilled: The nWER for the Cache LM system output was now 5.5% compared to 6.5% of the baseline system.

Language model interpolation

In this experiment, the LM of the baseline ASR system was interpolated with a small LM computed on the translations found in the MT n-best lists. The best system had a WER of 11.6%. The LM interpolation approach uses MT context information in the form of trigrams (and bigrams and monograms for back-off). The reduction of WER is relatively small when compared to the reductions obtained with the rescoring and cache LM approach. This can be explained by the limited input of MT context information.

Combination of ASR adaptation techniques

The proposed ASR improvement techniques apply different forms of MT knowledge with varying success. For this reason, we examined whether it is possible to further increase recognition accuracy by combining these techniques. Table 2 gives an overview of the WER accomplished with these different combinations together with the WER of previously described basic techniques.

MT adaptation techniques

To improve the Spanish-to-English MT system, we used the 150-best lists produced by the "Hypothesis Selection on Cache LM" approach. We tested two techniques and their combination. The results are summarized in Table 3.

As described above, the MT system consists of an LM, a TM, and a translation memory. Our MT system did not allow for the application of cache LMs without the need for modifications. Therefore, we limited ourselves to using LM interpolation for improving the target LM with the results from the ASR. For that purpose, we trained a small LM on the ASR 150-best lists and interpolated it with the original LM. As a result, the BLEU score increased to 53.4.

The TM computes phrase translation probabilities regardless of the word order. The word order of the translation is therefore appointed by the LM and translation memory. To retrain the MT system, the ASR 150-best lists were added several times to the original training data. The TM was then retrained, first with the translation memory fixed to the original training data and second with the translation memory computed over the complete training data. The best BLEU scores were 42.1 and 70.2, respectively.

In another step, we combined the above described systems for LM interpolation and retraining. The combination led to an improvement of the BLEU scores for the fixed and updated translation memory to 54.2 and 84.7, respectively.

Table 2. Comparison of ASR improvement techniques.

Technique	WER
Baseline ASR	12.6
LM Interpolation	11.6
Hypothesis Selection (on Baseline)	10.5
Cache LM	10.4
Cache and Interpolated LM	10.1
Hypothesis Selection on Cache and Interpolated LM	9.7
Hypothesis Selection on Cache LM	9.4

Table 3. Comparison of MT improvement techniques.

Technique	BLEU
Baseline MT	40.4
LM Interpolation	53.4
Updated Translation Memory	
– Retraining	70.2
– Combination	84.7
Fixed Translation Memory	
– Retraining	42.1
– Combination	54.2

Document-driven iterative MTEASR

After examining and optimizing the individual adaptation techniques separately, we tested our iterative system on Data Set II, first for the case that a written representation of the target language exists. For improving the ASR, the cache LM approach as well as the previously introduced combinations of techniques were examined. For the MT improvement, the combination of LM interpolation and retraining was chosen, on the one hand with a fixed translation memory and on the other hand with an updated memory. The motivation for this was that, although the MT system with the updated memory yielded a much higher performance, complementary MT knowledge that is valuable for further ASR improvement is lost by using it. An updated memory sees to it that, primarily, the ASR hypotheses added to the training data are selected as translation hypotheses. As a result, only a slightly changed ASR output of the preceding iteration is used for ASR improvement in the next iteration instead of new MT hypotheses. For improving the ASR component, the combination of rescoring and cache LM in iteration 0 and the combination of rescoring, cache LM, and interpolated LM in higher iterations yielded the best results. The better performance resulting from the additional use of LM interpolation after iteration 0 is due to the improved MT context information. For MT improvement, it turned out that it is better to work with a fixed translation memory. The final WER was 1% worse with the updated translation memory. No significant change in recognition accuracy was observed after one iteration. Figure 2 gives an overview on the components of our final iterative system design along with the respective performance values. With the iterative approach, we were able to reduce the WER of the English baseline ASR system from 20.4% to 13.1%.

Speech-driven iterative MTEASR

Experiments and results

Our final iterative design lifts the constraint of a textual representation of the source language by applying an ASR system to the source language speech. The same combinations of adaptation techniques as for the document-driven case yielded the best results. It was sufficient to improve the MT components just once within the iterative system design for gaining best results in speech recognition accuracy (for both involved ASR systems), just as in the document driven case. Figure 3 gives an overview of the components of our final speech driven iterative system design along with the respective performance values. The WER of the English ASR system was reduced from 20.4% to 14.3%. This is a relative reduction of 29.9%. The WER of the Spanish ASR of 17.2% was reduced by 20.9% relative to 13.6%. This smaller improvement in recognition accuracy compared to the improvement of the English ASR may be explained by the fact that Spanish is morphologically more complex than English. In iteration 0, the BLEU score of the Spanish-to-English MT system is 15.1%, relatively

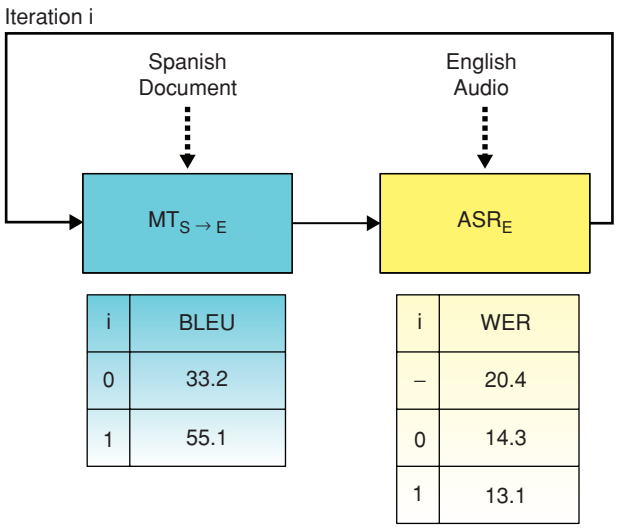


Fig. 2 Document driven iterative MTEASR

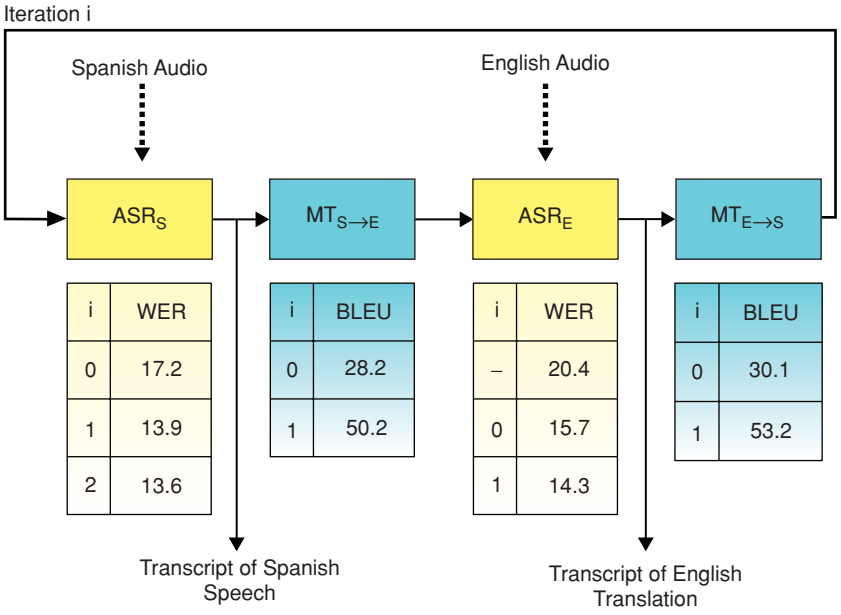


Fig. 3 Speech driven iterative MTEASR

worse than in the document-driven case. This results from the fact that the Spanish source sentences, which are used for translation, contain more noise due to recognition errors. In this context, it should be noted that this degradation in MT performance is of approximately the same magnitude as the WER of the Spanish input used for translation, i.e., it is of approximately the same magnitude as the WER of the Spanish baseline system. The degradation in MT performance leads to a smaller improvement of the English ASR system compared to the document-driven case. However, the loss in MT performance does not lead to a degradation in English speech recognition accuracy of the same magnitude; compared to the document-driven case, the WER of the English ASR system is only 9.8%, relatively higher. Figure 4 shows a detailed comparison of the performance of the English ASR system in the document driven and the speech driven case. Even though the gain in recognition accuracy is already remarkably high in both cases, without applying any iteration, a still significant gain in performance is to be observed in the first iteration.

Conclusions

In this article, we introduced an iterative system for improving speech recognition in the context of human-mediated translation scenarios. In contrast to related work conducted in this field, we included scenarios in which only spoken language representations are available. One key feature of our iterative system is that all involved system components, ASR as well as MT, are improved. Particularly in the context of a spoken source language representation, not only is the target language ASR automatically improved but so is the source language ASR. Using Spanish as the source language and English as the target language, we were able to reduce the WER of the English ASR by 35.8% when given a written-source language representation. Given a spoken-source language representation, we achieved a relative WER reduction of 29.9% for English and 20.9% for Spanish. This iterative system design also allows for the incorporation of knowledge provided by not just one audio stream in another language but by many. Only minimal modifications of the applied adaptation techniques would be necessary for such a scenario. The adaptation of the cache LM approach as well as the LM interpolation (for ASR and MT improvement) and MT retraining can be done by including all MT/ASR n-best lists of the preceding MT/ASR systems in the iterative cycle. The ASR rescoring algorithm can be extended to allow for several TM scores provided by several MT systems with different target languages.

Read more about it

- M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Speech translation enhanced automatic speech recognition," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Puerto Rico, Dec. 2005, pp. 121–126.
- M. Paulik, S. Stüker, C. Fügen, T. Schultz, T. Schaaf, and A. Waibel, "Document driven machine translation enhanced ASR," in *Proc. 9th European Conf. Speech Commun. and Technol.*,

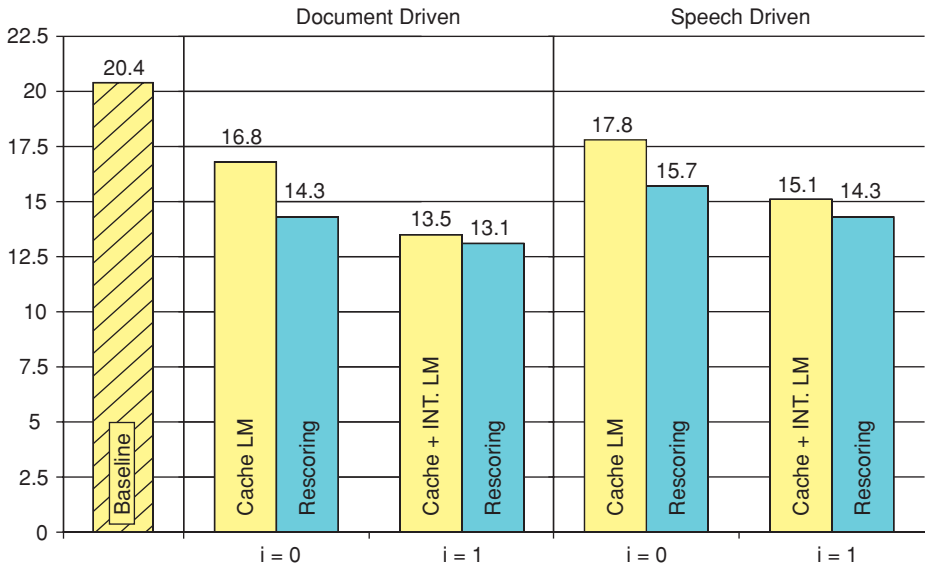


Fig. 4 Detailed comparison of the document- and speech-driven case

Lisbon, Portugal, Sep. 2005, 2,261–2,264.

• M. Dymetman, J. Brousseaux, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, "Towards an automatic dictation system for translators: The transtalk project," in *Proc. ICSLP*, Yokohama, Japan, 1994.

• P. Brown, S. Della Pietra, S. Chen, V. Della Pietra, S. Kehler, and R. Mercer, "Automatic speech recognition in machine aided translation," in *Comput. Speech and Language*, vol. 8, no. 3, pp. 177-187, 1994.

• J. Brousseaux, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon, "French speech recognition in an automatic dictation system for translators: The transtalk project," in *Proc. Eurospeech*, Madrid, Spain, 1995, 177–187.

• Y. Ludovik and R. Zacharski, "MT and topic-based techniques to enhance speech recognition systems for professional translators," in *Proc. CoLing,* Saarbrücken, Germany, 2000, pp.193–196.

• H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment" in *Proc. ASRU*, Madonna di Campiglio, Italy, 2001, pp. 1061–1065.

• F. Metze, Q. Jin, C. Fügen, K. Laskowski, Y. Pan, and T. Schultz, "Issues in meeting transcription—The ISL meeting transcription system," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 1703–1712.

• S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel, "The ISL statistical machine translation system for spoken language translation," in *Proc. IWSLT*, Kyoto, Japan, 2004, pp. 65–72.

About the authors

Matthias Paulik (paulik@ira.uka.de) received his degree in computer science from Universität Karlsruhe, Germany, in May 2005. He is a candidate in the doctoral program of the Department of Computer Science at Universität Karlsruhe where he works in the interACT Research Laboratories, Carnegie Mellon, USA. The focus of his research is the development of speech translation systems by learning from human interpreters. He is a Graduate Student Member of the IEEE.

Sebastian Stüker (stueker@ira.uka.de) received his diploma degree in computer science from Universität Karlsruhe, Germany, in May 2003. He is a member of the doctoral program of the Department of Computer Science at Universität Karlsruhe where he holds a fully funded research position. He is a member of the IEEE and vice chair of the IEEE Karlsruhe Student Branch.

Christian Fügen (fuegen@ira.uka.de) received his diploma degree in computer science from Universität Karlsruhe, Germany, in December 1999. He is a member of the doctoral program of the Department of computer Science at Universität Karlsruhe where he holds a fully funded research position. He is a member of the IEEE and ISCA.

Tanja Schultz (tanja@cs.cmu.edu) received her Ph.D. and master's degree in computer science from University Karlsruhe, Germany, in 2000 and 1995 respectively. She received a German master's in mathematics, sports, and education science from the University of Heidelberg, Germany, in 1990. She is a research assistant professor at Carnegie Mellon. Her research centers on multilingual speech processing as well as speech translation systems.

Alex Waibel is a professor at the School of Computer Science at Carnegie Mellon University and at the Computer Science Department of Karlsruhe University, Germany. He received his B.S. degree from the Massachusetts Institute of Technology in 1979, his M.S. (electrical engineering and computer science) in 1980, and his Ph.D. (computer science) from Carnegie Mellon University in 1986. He is the director of InterACT, the International Center of Advanced Communication Technologies at both universities. He holds joint appointments in the Language Technologies Institute, the Human Computer Interaction Institute, the Computer Science Department and the Robotics Institute. He was one of the founding members of the Human Computer Interaction Institute at Carnegie Mellon and serves on the steering committee.