# Incremental Adaptation of Speech-to-Speech Translation

**Nguyen Bach, Roger Hsiao, Matthias Eck, Paisarn Charoenpornsawat, Stephan Vogel, Tanja Schultz, Ian Lane, Alex Waibel and Alan W. Black**
InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nbach, wrhsiao, matteck, paisarn, stephan.vogel, tanja, ianlane, ahw, awb}@cs.cmu.edu

## Abstract

In building practical two-way speech-to-speech translation systems the end user will always wish to use the system in an environment different from the original training data. As with all speech systems, it is important to allow the system to adapt to the actual usage situations. This paper investigates how a speech-to-speech translation system can adapt day-to-day from collected data on day one to improve performance on day two. The platform is the CMU Iraqi-English portable two-way speech-to-speech system as developed under the DARPA TransTac program. We show how machine translation, speech recognition and overall system performance can be improved on day 2 after adapting from day 1 in both a supervised and unsupervised way.

## 1 Introduction

As speech-to-speech translation systems move from the laboratory into field deployment, we quickly see that mismatch in training data with field use can degrade the performance of the system. Retraining based on field usage is a common technique used in all speech systems to improve performance. In the case of speech-to-speech translation we would particularly like to be able to adapt the system based on its usage automatically without having to ship data back to the laboratory for retraining. This paper investigates the scenario of a two-day event. We wish to improve the system for the second day based on the data collected on the first day.

Our system is designed for eyes-free use and hence provides no graphical user interface. This allows the user to concentrate on his surrounding environment during an operation. The system only provides audio control and feedback. Additionally the system operates on a push-to-talk method. Previously the system (Hsiao et al., 2006; Bach et al., 2007) needed 2 buttons to operate, one for the English speaker and the other one for the Iraqi speaker.

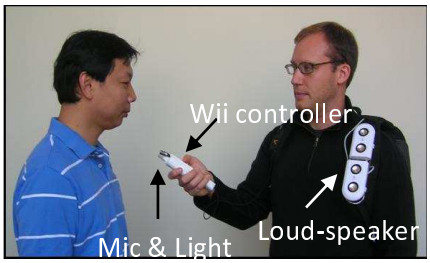


Figure 1: The users interact with the system

To make the system easier and faster to use, we propose to use a single button which can be controlled by the English speaker. We mounted a microphone and a Wii remote controller together as shown in 1.

Since the Wii controller has an accelerometer which can be used to detect the orientation of the controller, this feature can be applied to identify who is speaking. When the English speaker points towards himself, the system will switch to English-Iraqi translation. However, when the Wii is pointed towards somebody else, the system will switch to Iraqi-English translation. In addition, we attach a light on the Wii controller providing visual feedback. This can inform an Iraqi speaker when to start speaking. The overall system is composed of five major components: two automatic speech recognition (ASR) systems, a bidirectional statistical machine translation (SMT) system and two text-to-speech (TTS) systems.

## 2 Data Scenario

The standard data that is available for the TransTac project was collected by recording human interpreter mediated dialogs between war fighters and Iraqi native speakers in various scenarios. The dialog partners were aware that the data was being collected for training machine based translation devices, but would often talk directly to the human interpreter rather than pretending it was an automatic device. This means that the dialog



*Proceedings of NAACL HLT 2009: Short Papers*, pages 149–152,
Boulder, Colorado, June 2009. 

partners soon ignored the recording equipment and used a mostly natural language, using informal pronunciation and longer sentences with more disfluencies than we find in machine mediated translation dialogs.

Most users mismatch their language when they communicate using an automatic speech-to-speech translation system. They often switch to a clearer pronunciation and use shorter and simpler sentences with less disfluency. This change could have a significant impact on speech recognition and machine translation performance if a system was originally trained on data from the interpreter mediated dialogs.

For this reason, additional data was collected during the TransTac meeting in June of 2008. This data was collected with dialog partners using the speech-to-speech translation systems from 4 developer participants in the TransTac program. The dialog partners were given a description of the specific scenario in form of a rough script and had to speak their sentences into the translation systems. The dialog partners were not asked to actually react to the potentially incorrect translations but just followed the script, ignoring the output of the translation system. This has the effect that the dialog partners are no longer talking to a human interpreter, but to a machine, pressing push-to-talk buttons etc. and will change their speech patterns accordingly.

The data was collected over two days, with around 2 hours of actual speech per day. This data was transcribed and translated, resulting in 864 and 824 utterance pairs on day 1 and 2, respectively.

## 3 ASR LM Adaptation

This section describes the Iraqi ASR system and how we perform LM adaptation on the day 1 data to improve ASR performance on day 2. The CMU Iraqi ASR system is trained with around 350 hours of audio data collected under the TransTac program. The acoustic model is speaker independent but incremental unsupervised MLLR adaptation is performed to improve recognition. The acoustic model has 6000 codebooks and each codebook has at most 64 Gaussian mixtures determined by merge-and-split training. Semi-tied covariance and boosted MMI discriminative training is performed to improve the model (Povey et al., 2009). The features for the acoustic model is the standard 39-dimension MFCC and we concatenate adjacent 15 frames and perform LDA to reduce the dimension to 42 for the final feature vectors. The language model of the ASR system is a trigram LM trained on the audio transcripts with around three million words with Kneser-Ney smoothing (Stolcke, 2002).

To perform LM adaptation for the ASR system, we use the ASR hypotheses from day 1 to build a LM. This LM is then interpolated with the original trigram LM to produce an adapted LM for day 2. We also evaluate the effect of having transcribers provide accurate transcription references for day 1 data, and see how it may improve the performance on day 2. We compare unigram, bigram and trigram LMs for adaptation. Since the amount of day 1 data is much smaller than the whole training set and we do not assume transcription of day 1 is always available, the interpolation weight is chosen of be 0.9 for the original trigram LM and 0.1 for the new LM built from the day 1 data. The WER of baseline ASR system on day 1 is 32.0%.

| Base | 1-g hypo | 2-g hypo | 3-g hypo | 1-g ref | 2-g ref | 3-g ref |
|---|---|---|---|---|---|---|
| 31.3 | 30.9 | 31.2 | 31.1 | 30.6 | 30.5 | 30.4 |

Table 1: Iraqi ASR's WER on day 2 using different adaptation schemes for day 1 data

The results in Table 1 show that the ASR benefits from LM adaptation. Adapting day 1 data can slightly improve the performance of day 2. The improvement is larger when day 1 transcript is available which is expected. The result also shows that the unigram LM is the most robust model for adaptation as it works reasonably well when transcripts are not available, whereas bigram and trigram LM are more sensitive to the ASR errors made on day 1.

| | Day 1 | Day 2 |
|---|---|---|
| No ASR adaptation | 29.39 | 27.41 |
| Unsupervised ASR adaptation | 31.55 | 27.66 |
| Supervised ASR adaptation | 32.19 | 27.65 |

Table 2: Impact of ASR adaptation to SMT

Table 2 shows the impact of ASR adaptation on the performance of the translation system in BLEU (Papineni et al., 2002). In these experiments we only performed adaptation on ASR and still using the baseline SMT component. There is no obvious difference between unsupervised and supervised ASR adaptation on performance of SMT on day 2. However, we can see that the difference in WER on day 2 of unsupervised and supervised ASR adaptation is relatively small.

## 4 SMT Adaptation

The Iraqi-English SMT system is trained with around 650K sentence pairs collected under the TransTac program. We used PESA phrase extraction (Vogel, 2005) and a suffix array language model (Zhang and Vogel, 2005). To adapt SMT components one approach is to optimize LM interpolation weights by minimizing perplexity of the 1-best translation output (Bulyko et al., 2007). Related work including (Eck et al., 2004) attempts to use information retrieval to select training sentences similar to those in the test set. To adapt the SMT components we use a domain-specific LM on top of the background

language models. This approach is similar to the work in (Chen et al., 2008). sThe adaptation framework is 1) create a domain-specific LM via an n-best list of day 1 machine translation hypothesis, or day 1 translation references; 2) re-tune the translation system on day 1 via minimum error rate training (MERT) (Venugopal and Vogel, 2005).

| Use | | Day 1 | Day 2 |
|---|---|---|---|
| | Baseline | 29.39 | **27.41** |
| 500 Best MT Hypos | 1gramLM | 29.18 | 27.23 |
| | 2gramLM | 29.53 | 27.50 |
| | 3gramLM | 29.36 | 27.23 |

Table 3: Performance in BLEU of unsupervised adaptation.

The first question we would like to address is whether our adaptation obtains improvements via an unsupervised manner. We take day 1 baseline ASR hypothesis and use the baseline SMT to get the MT hypothesis and a 500-best list. We train a domain LM using the 500-best list and use the MT hypotheses as the reference in MERT. We treat day 1 as a development set and day 2 as an unseen test set. In Table 3 we compare the performance of four systems: the baseline which does not have any adaptation steps; and 3 adapted systems using unigram, bigram and trigram LMs build from 500-best MT hypotheses.

| Use | | Day 1 | Day 2 |
|---|---|---|---|
| | Baseline (no tune) | 29.39 | 27.41 |
| | Baseline (tune) | 29.49 | 27.30 |
| 500 Best MT Hypos | 1gramLM | 30.27 | **28.29** |
| | 2gramLM | 30.39 | **28.30** |
| | 3gramLM | 28.36 | 24.64 |
| MT Ref | 1gramLM MT Ref | 30.53 | **28.35** |

Table 4: Performance in BLEU of supervised adaptation.

Experimental results from unsupervised adaptation did not show consistent improvements but suggest we may obtain gains via supervised adaptation. In supervised adaptation, we assume we have day 1 translation references. The references are used in MERT. In Table 4 we show performances of two additional systems which are the baseline system without adaptation but tuned toward day 1, and the adapted system which used day 1 translation references to train a unigram LM (1gramLM MT Ref). The unigram and bigram LMs from 500-best and unigram LM from MT day 1 references perform relatively similar on day 2. Using a trigram 500-best LM returned a large degradation and this LM is sensitive to the translation errors on day1

## 5 Joint Adaptation

In Sections 3 and 4 we saw that individual adaptation helps ASR to reduce WER and SMT to increase BLEU score. The next step in validating the adaptation framework was to check if the joint adaptation of ASR and SMT on day 1 data will lead to improvements on day 2. Table 5 shows the combination of ASR and SMT adaptation methods. Improvements are obtained by using both ASR and SMT adaptation. Joint adaptation consistently gained more than one BLEU point improvement on day 2. Our best system is unsupervised ASR adaptation via 1gramLM of ASR day 1 transcription coupled with supervised SMT adaptation via 1gramLM of day 1 translation references. An interesting result is that to have a better result on day 2 our approach only requires translation references on day 1. We selected 1gramLM of 500-best MT hypotheses to conduct the experiments since there is no significant difference between 1gramLM and 2gramLM on day 2 as showed in Table 3.

| ASR | SMT | Day 1 | Day 2 |
|---|---|---|---|
| No adaptation | No adaptation | 29.39 | 27.41 |
| Unsupervised ASR adaptation with 1gramLM ASR hypo | 1gramLM 500-Best MT Hypo | 32.07 | 28.65 |
| | 1gramLM MT Ref | 31.76 | **28.83** |
| Supervised ASR adaptation with 1gramLM transcription | 1gramLM 500-Best MT Hypo | 32.48 | 28.59 |
| | 1gramLM MT Ref | 32.68 | 28.60 |

Table 5: Performance in BLEU of joint adaptation.

## 6 Selective Adaptation

The previous results indicate that we require human translation references on day 1 data to get improved performance on day 2. However, our goal is to make a better system on day 2 but try to minimize human efforts on day 1. Therefore, we raise two questions: 1) Can we still obtain improvements by not using all of day 1 data? and 2) Can we obtain more improvements?

To answer these questions we performed oracle experiments when we take the translation hypotheses on day 1 of the baseline SMT and compare them with translation references, then select sentences which have BLEU scores higher than a threshold. The subset of day 1 sentences is used to perform supervised adaptation in a similar way showed in section 5. These experiments also simulate the situation when we have a perfect confidence score for machine translation hypothesis selection. Table 6 shows results when we use various portions of day 1 to perform adaptation. By using day 1 sentences which have smoothed sentence BLEU scores higher than 10 or 20 we have very close performance with adaptation by using all day 1 data. The results also show that by using 416 sentences which have sentence BLEU score higher than 40 on day 1, our adapted translation components outperform the baseline. Performance starts degrading after 50. Experimental results lead to the answer for question 1) that

by using less day 1 data our adapted translation components still obtain improvements compare with the baseline, and 2) we did not see that using less data will lead us to a better performance compare with using all day 1 data.

| | No. sents | Day 1 | Day 2 |
|---|---|---|---|
| Baseline | | 29.39 | 27.41 |
| $\geq 0$ | 864 | 30.27 | 28.29 |
| $\geq 10$ | 797 | 31.15 | **28.27** |
| $\geq 20$ | 747 | 30.81 | **28.24** |
| $\geq 30$ | 585 | 30.04 | **27.71** |
| $\geq 40$ | 416 | 29.72 | **27.65** |
| $\geq 50$ | 296 | 30.06 | 27.04 |
| Correct | 98 | 29.18 | 27.19 |

Table 6: Performance in BLEU of selective adaptation

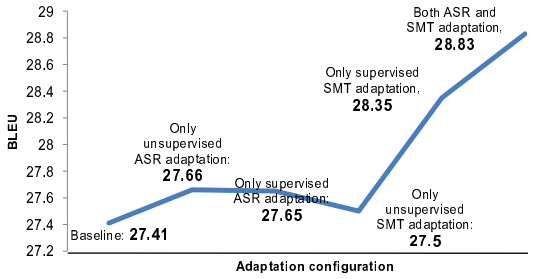


Figure 2: Summarization of adaptation performances

## 7 Conclusions

This work clearly shows that improvement is possible using collected data for adaptation. The overall picture is shown in Figure 2. However this result is only based on one such data set, it would be useful to do such adaptation over multiple days. The best results however still require producing translation references, notably ASR transcriptions do not seem to help, but may still be required in the process of generating translation references. We wish to further investigate automatic adaptation based on implicit confidence scores, or even active participation of the user e.g. by marking bad utterance which could be excluded from the adaptation.

## Acknowledgments

This work is in part supported by the US DARPA under the TransTac (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. We would also like to thank Cepstral LLC and Mobile Technologies LLC, for support of some of the lower level software components.

## References

Nguyen Bach, Matthias Eck, Paisarn Charoenpornsawat, Thilo Khler, Sebastian Stker, ThuyLinh Nguyen, Roger Hsiao, Alex Waibel, Stephan Vogel, Tanja Schultz, and Alan Black. 2007. The CMU TransTac 2007 Eyes-free and Hands-free Two-way Speech-to-Speech Translation System. In *Proc. of the International Workshop on Spoken Language Translation*, Trento, Italy.

Ivan Bulyko, Spyros Matsoukas, Richard Schwartz, Long Nguyen, and John Makhoul. 2007. Language Model Adaptation in Machine Translation from Speech. In *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, Honolulu, Hawaii, USA.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting n-best hypotheses for smt self-enhancement. In *Proceedings of ACL-08: HLT, Short Papers*, pages 157–160, Columbus, Ohio, USA, June.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2004. Language model adaptation for statistical machine translation based on information retrieval. In *Proc. LREC'04*, Lisbon, Portugal.

Roger Hsiao, Ashish Venugopal, Thilo Kohler, Ying Zhang, Paisarn Charoenpornsawat, Andreas Zollmann, Stephan Vogel, Alan W Black, Tanja Schultz, and Alex Waibel. 2006. Optimizing Components for Handheld Two-way Speech Translation for an English-Iraqi Arabic System. In *Proc. of Interspeech*, Pittsburgh, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318, Philadelphia, PA, July.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhu vana Ramabhadran, George Saon, and Karthik Visweswariah. 2009. Boosted MMI for model and feature-space discriminative training. In *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, USA.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Ashish Venugopal and Stephan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of EAMT-05*, Budapest, Hungary.

Stephan Vogel. 2005. Pesa: Phrase pair extraction as sentence splitting. In *Proc. of MT SUMMIT X*, Phuket, Thailand.

Ying Zhang and Stephan Vogel. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of EAMT'05*, Budapest, Hungary, May. The European Association for Machine Translation.