# Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems

*Shirin Saleem, Szu-Chen Jou, Stephan Vogel and Tanja Schultz*

Interactive Systems Lab
Language Technologies Institute
Carnegie Mellon University, U.S.A.
{ssaleem,scjou,stephan.vogel,tanja}@cs.cmu.edu

## Abstract

In this paper we present first experiments towards a tighter coupling between Automatic Speech Recognition (ASR) and Statistical Machine Translation (SMT) to improve the overall performance of our speech translation system. In coventional speech translation systems, the recognizer outputs a single hypothesis which is then translated by the SMT system. This approach has the limitation of being largely dependent on the word error rate of the first best hypothesis. The word error rate is typically lowered by generating many alternative hypotheses in the form of a word lattice. The information in the word lattice and the scores from the recognizer can be used by the translation system to obtain better performance. In our experiments, by switching from the single best hypotheses to word lattices as the interface between ASR and SMT, and by introducing weighted acoustic scores in the translation system, the overall performance was increased by 16.22%.

## 1. Introduction

Speech translation has made significant advances over the last few years moving from highly domain restricted and planned tasks to conversational speech translation in less limited domains. Due to the peculiarities of spoken language, an effective solution to speech translation cannot be expected to be a mere sequential connection of Automatic Speech Recognition (ASR) and Machine Translation (MT) components. According to Ringger [1], the coupling between ASR and MT can be characterized by three orthogonal dimensions: the (a) complexity of the search algorithm, e.g. extent to which the language model used for parsing, the (b) incrementality, that indicates whether the entire recognizer hypothesis is processed by components at the next level at a time, or incrementally, and (c) tightness, that describes if ASR and MT closely interact while searching for a solution (tight), exchange some information (semi-tight), or do not interact at all (loose). The benefits and drawbacks have been widely discussed along aspects such as modularity, scalability, and complexity of systems [1].

There are several systems for spoken language translation that use different coupling strategies. MASTOR [2] is IBM's highly trainable loosely coupled speech to speech translation system targeting conversational spoken language translation between English and Mandarin Chinese for limited domains. Other examples of loosely coupled systems are Diplomat [3], a speech to speech translation system developed at CMU that can easily adapt to new languages, and ATR-MATRIX [4], a system that performs translation of spontaneous Japanese speech into English in nearly real time. Examples for tightly coupled systems using finite state transducers are EuTrans [5] developed at UPV and AT&T's Transnizer [6]. The EuTrans architecture is based on the ATROS [7] engine. The use of finite state models allows these systems to obtain translation synchronously with the recognition process. In AT&T's approach multimodal parsing, understanding, and integration are achieved using a finite-state model [6]. There are other systems that use Interlingua design to translate spontaneously spoken dialogues such are JANUS [8], and Nespole! [9]. In [10], Ney discusses the coupling between recognition and translation using the methods of local averaging approximation and monotone alignments.

In order to support de-coupled implementation and improvements of ASR and MT components, as well as scalable systems that allow complex translation tasks, our system structure is designed as a complex, non-incremental, loose coupling of ASR and MT. So far this coupling was realized via the single best hypothesis generated by the ASR. This however, has the limitation that the MT unit is largely dependent on the word error rate of the single best hypothesis. Typically the word error rate can be decreased by generating many alternatives in the form of n-best lists or word lattices [11]. Therefore, we expect the translation performance to also benefit by getting access to these alternatives. In order to achieve an overall optimal output, it is necessary to also incorporate the ASR scores into the selection process of the MT. Otherwise the MT component might favor strings which are easy to translate but have a high word error rate. The following sections describe experiments in which word lattices are used as the interface between ASR and MT. The acoustic scores of the words in each path of the lattice are also added to the total translation score.

## 2. System Description

### 2.1. Data

The experiments were done on German to English translation. The data used for training and testing was originally collected as part of the Nespole! speech translation project [9]. It consists of spontaneous dialogs spoken in four European languages, including German and English, in the travel and tourism domain between a tourist information service provider and a customer who wants to organize a trip.

The audio data was recorded using a close-speaking microphone and sampled at 16 KHz at a resolution of 16 bit. The acoustic models of the ASR component (see section 2.2) were trained using 62 hours of recorded speech; a corpus of 650K tokens and 10.5K types was applied for the language model training. The MT component (see section 2.3) is trained on German–English parallel text for both translation model and language model. The data contains about 15K words for each language. To make the language model more robust, it is interpolated with the Verbmobil English text data which contains about 500K words in the travel domain.

The test data used was part of the Nespole! Showcase-1 evaluation data dated November 2001. Two development sets of 52 utterances, and 23 utterances respectively were used in the experiments. A test set of 70 utterances in German by a single male client forms the basis for the results reported here. A description of these test sets can be found in Table 1.

*Table 1:* Test Data Statistics

| | Utterances | Tokens | Types |
|---|---|---|---|
| Dev set 1 | 23 | 250 | 87 |
| Dev set 2 | 52 | 342 | 157 |
| Test set | 70 | 438 | 184 |

### 2.2. Speech Recognizer

In our experiments, we used the JANUS speech recognition toolkit (JRTk) with the Ibis decoder [12]. It achieves 26.71% word error rate in 1.3x real time on the above given test set

In JRTk, a word lattice is represented as a directed graph where the nodes are associated with words and the links represent the possible succession of words in the different hypotheses. The acoustic word scores are stored in the links rather than in the nodes of the lattice. JRTk has lattice related functions explicitly for beam width pruning, filler word removal and lattice output. Beam width pruning can control the word branch factor to reduce lattice density to some expected range so that lattices provide different degree of recovery with different densities. Filler words are non speech events, e.g. lip smack, or noise hypothesized by the speech recognizer. Since filler words are not used in the translation component, they are removed from the recognizer output.

### 2.3 Machine Translation System

The translation model used is the CMU Statistical Machine Translation toolkit (CMU-SMT) [13]. It contains a lexical transducer, phrase transducer and a class based transducer. The lexical transducer is a one-to-one lexicon mapper, the phrase transducer is a many-to-one lexicon mapper, and the class based transducer maps the word classes such as weekdays, numbers, etc. The language model is n-gram based and up to trigrams are used.

The input to the SMT system is a one best sentence or a lattice. The one best sentence if inputted is converted into a single path lattice first. The lattice that is fed into the SMT system has a different format from that of the JRTk lattice. Here, the nodes of the lattice contain the acoustic scores, and the words are attached to the edges. For each edge (source word), the transducers are applied to create corresponding target word edges between the same vertices on the same word lattice. Then the enlarged word lattice represents a search space including all the source and target word edges. The next step is to find a best path in the word lattice with path scores accumulated with transducer scores, target language model scores, and acoustic scores.

### 2.4. Coupling between JRTk and CMU-SMT system

The JRTk speech recognizer generates an n-best list or a word lattice based on the probability of each word sequence according to the acoustic and language model. Translating an n-best list one sentence at a time leads to a linear growth in translation time. To make translation more efficient, word lattices are used instead. The lattice can be pruned to different densities. Density is defined as number of words in the lattice divided by the number of words in the transliteration. However, the lattice may contain paths that are worse than any of the original hypotheses in the n-best list. Thus there is a need to combine the scores from recognition and translation. The acoustic model scores can be propagated with the individual words in the lattice.

## 3. Experiments

For evaluation in our experiments we apply the modified BLEU automatic evaluation metric as proposed by IBM [14]. Two human reference translations of the test set were used in the calculation of the BLEU score.

### 3.1. Impact of lattice density on Word Error Rate

Lattice word error rate is defined as the minimal word error rate of all possible paths through the lattice. Figure 1 shows that the word error rate decreases with growing lattice density. In our case the word error rate of the first best hypothesis (density 1) is reduced by 41.53% by moving to a lattice of density 7.

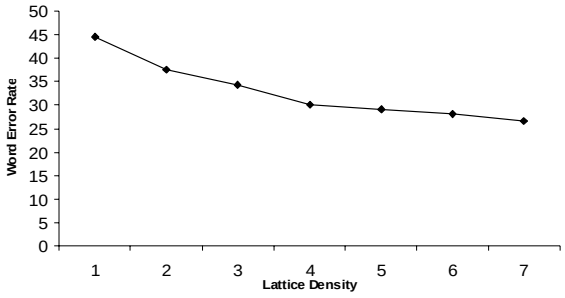


*Figure 1:* Word Error Rate versus Lattice Density

### 3.2. Lattices versus first best hypothesis as input to the MT unit

In our first experiment, we compared the BLEU scores of the hypotheses of the MT unit when translating on JRTk word lattices of different densities. We found that the highest BLEU score of 0.1683 was achieved for the single best hypothesis. In this experiment, neither acoustic scores, nor

language model scores of the paths in the JRTk lattice were taken into account. Only the advantage of lattice topology of different densities as input to the SMT system was considered. This resulted in the MT unit translating paths through the lattice which performed worse that the one best case. The one best hypothesis, on the other hand is formed by applying the acoustic scores and source language model scores by the speech recognizer. This trade-off makes the word lattice method worse. In the next experiment, we add the weighted acoustic scores to the translation scores and see an improvement in the BLEU score.

### 3.3. Incorporating acoustic scores in the MT unit

In this experiment the weighted acoustic scores from the speech recognizer were added to the total translation score. The sum of the acoustic scores of the chosen word sequence was weighted with factors ranging from 0.01 to 0.29, and the BLEU score of the hypotheses of the MT unit was calculated. This was done for lattices of different densities. We noted that there is an improvement of the BLEU score over that of the one-best hypothesis. But there is no smooth transition of scores with increasing density. The addition of acoustic scores alone does not guarantee that the optimal path through the lattice is chosen by the MT unit. This calls for the need to include the source language model scores as well. Figure 2 shows the results of the experiments for acoustic score weights of 0, 0.02 and 0.28. The improvement over the baseline BLEU score for the one best hypothesis for a lattice of density 3, with acoustic scores weighted by factor of 0.28 is 7.3%.

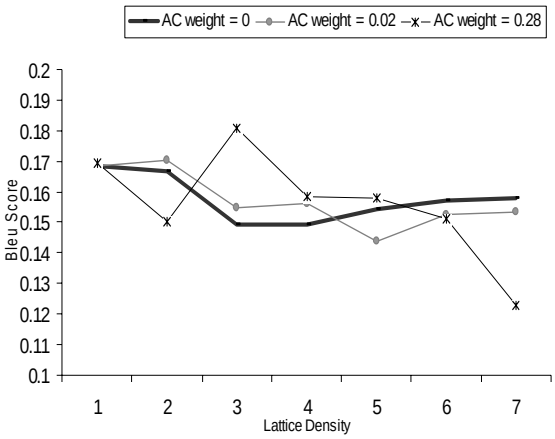


*Figure 2:* Modified Bleu score versus Lattice Density

### 3.4. Impact of sentence length on BLEU score

This experiment was carried out to observe the effect of sentence length on the BLEU score. First, the utterances were statistically grouped into short, medium, and long test sets based on the number of words in the audio transliteration. Then the BLEU score for increasing lattice densities was calculated separately for each of these test sets. For an acoustic score weight of 0.01, Figure 3 shows the variation of the BLEU score with lattice density for the short, medium and long test sets. We see that the improvement over the first best hypothesis by switching to lattices was greatest in the case of the long utterances.

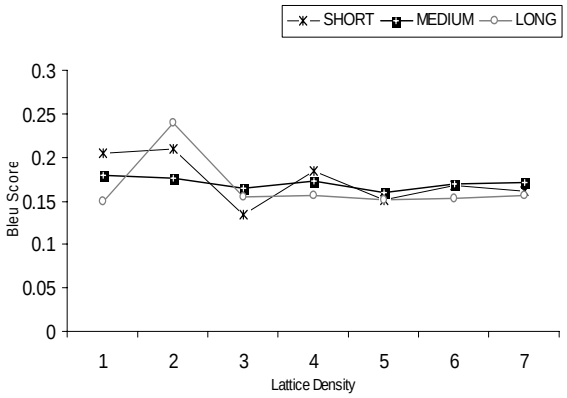


*Figure 3:* Modified BLEU score versus lattice density for utterances of different lengths

### 3.5. Tuning the system for optimal parameters

Experiments were carried out on the development sets to find the optimal lattice density and acoustic score weight individually for short, medium and long utterances. The optimized parameters were then applied to the test set to get an improvement in BLEU score of 13.23%. When the parameters are tuned for the test set to get the best performance , the improvement is 16.22% over the baseline BLEU of the first best hypothesis. Table 2 shows the best parameters for the test set.

*Table 2:* Optimal density and acoustic score weight based on utterance length

| | Number of Words | Optimal AC Weight | Optimal Density |
|---|---|---|---|
| Short | 1-5 | 0.08 | 4 |
| Medium | 6-10 | 0.22 | 3 |
| Long | 10-23 | 0.01 | 2 |

## 4. Conclusions and Future Work

The experiments described in this paper show that using word lattices as the interface between ASR and MT does improve translation performance when the weighted acoustic scores are incorporated into the MT unit. However, the limitation of not having the source language model scores as well is evident. Unlike the acoustic score, the source language model scores cannot be propagated from the recognizer to the MT unit with the words of the lattice. The language model score depends on the word history and each word can have multiple histories. So attaching the language model scores to the words of the lattice would blow up the size of the lattice. The source language model would thus, have to be integrated into the SMT system. This would be the next step in our implementation

## 5. References


[1] E. K. Ringger, “A robust loose coupling for speech recognition and natural language understanding”, *Technical Report 592*., University of Rochester Computer Science Department, 1995.

[2] F. H. Liu, L. Gu, Y. Gao, M. Picheny, “Use of Statistical N-Gram Models in Natural Language Generation for Machine Translation”, *Proceedings of ICASSP*, pp. 636-639, April 2003.

[3] A. W. Black, R. D. Brown, R. Frederking, K. Lenzo, J. Moody, A. Rudnicky, R. Singh, E. Steinbrecher. “Rapid Development of Speech-to-Speech Translation Systems”, *Proceedings of ICSLP*, pp 1709-1712, Sept 2002.

[4] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, S. Yamamoto, “A Japanese-to-English Speech Translation System:ATR-MATRIX”, *Proceedings of ICSLP*, pp: 2779-2782, Dec 1998.

[5] M. Pastor, A. Sanchis, F. Casacuberta, E. Vidal,“EuTrans: a Speech-to-Speech Translator Prototype”, *Proceedings of Eurospeech,* pp: 2385-2389, Sept 2001.

[6] S. Bangalore, G. Riccardi, "A Finite-State Approach to Machine Translation", *Proceedings of NAACL*, May 2001.

[7] D. Llorens, F. Casacuberta, E. Segarra, J.A. S´anchez, P. Aibar, “Acoustical and syntactical modeling in ATROS system,” *Proceedings of ICASSP*, pp. 641–644, May 1999.

[8] M. Woszczyna, N. Coccaro, A. Eisele, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. P. Rose, T. Sloboda, M.Tomita, J. Tsutsumi, N. Aoki-Waibel, A. Waibel, W. Ward, ″Recent Advances in JANUS: A Speech Translation System", *Proceedings of Eurospeech*, pp: 1295-1298, Sept 1993.

[9] F. Metze, C. Langley, A. Lavie, J. McDonough, H. Soltau, A. Waibel, S. Burger, K. Laskowski, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei., "The NESPOLE! Speech-to-Speech Translation System", *Proceedings of HLT*, 2003

[10] H. Ney, “Speech Translation: Coupling of Recognition and Translation“, *Proceedings of ICASSP*, pp. 517-520, May 1999.

[11] P. Woodland, C. Leggetter, J. Odell, V. Valtchev, and S. J. Young, "The Development of the 1994 HTK Large Vocabulary Speech Recognition System", Proceedings of the ARPA Spoken Language Systems Technology Workshop. ARPA, ARPA**,** January 1995.

[12] H. Soltau, F. Metze, C. Fgen, A. Waibel, "A one pass-decoder based on polymorphic linguistic context assignment", *Proceedings of ASRU*, Dec 2001.

[13] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, A. Waibel, "The CMU Statistical Machine Translation System", *Proceedings of MT-Summit IX*, LA. Sep 2003.

[14] K. Papineni, S. Roukos, T. Ward, W. Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proceedings of ACL*, pp. 311-318, July 2002.