Phonetic Representation-Based Speech Translation

Jie Jiang,[†] **Zeeshan Ahmed,**[‡] **Julie Carson-Berndsen,**[‡] **Peter Cahill,**[‡] and **Andy Way**[†]

[†]Applied Language Solutions, Delph, Greater Manchester, United Kingdom [*]
{jie.jiang, andy.way}@appliedlanguage.com

[‡]Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
Zeeshan.ahmed@ucdconnect.ie,{julie.berndsen,peter.cahill}@ucd.ie

Abstract

This paper explores a tight coupling of Automatic Speech Recognition (ASR) and Machine Translation (MT) for speech translation with information sharing on the phone-level. Our novel approach allows MT to access fine-grained phonetic information from ASR, as a methodology for facilitating speech translation. Specifically, Phrase-based Statistical MT (PBSMT) models are adapted to work on source language phones, and with a configuration of a source-language phoneme-to-grapheme component, source-language phones are translated into target-language words. Furthermore, to take advantage of source-side phonetic confusion information from the speech recogniser, phone confusion networks are constructed from the phonetic confusion matrix and are used as SMT inputs to boost translation quality. Experiments are carried out on IWSLT English–Chinese translation task, and significant improvements (1.27 absolute and 4.29% relatively BLEU points) are obtained by using phone confusion networks over the baseline PBSMT system.

1 Introduction

With the recent progress of automatic speech recognition (ASR) and machine translation (MT), speech translation has become an active research domain. The most commonly used speech translation model is the cascaded approach, which treats ASR and MT as black boxes, and use words as the basic unit for information sharing between these two components. In this architecture, source language speech is fed into an ASR module to obtain recognition results in 1-best, n-best (Zhang et al., 2004), word lattice (Matusov et al., 2005b), confusion network (Bertoldi et al., 2008a) formats, then the recognised outputs are translated with MT modules into a target language. MT outputs are finally used in a Speech Synthesis (SS) module for target-side human interaction. This approach is straight forward and can easily benefit from the improvements of any of these components.

However, there are still some limitations of this approach:

- Recognition errors introduced by the ASR module propagate into the following modules, and sometimes it is difficult to recover based on word-level information.
- It is difficult to tune three modules together since they are built from different corpora, under the assumption that the source-side ASR vocabulary is coherent with the MT vocabulary. However, each module is typically trained from independent, mismatched corpora as not many have been designed specifically for the speech translation task.
- Information sharing between the ASR, MT, and SS modules is weak, as the MT module generally processes at the word-level, source-side acoustic details are lost (for example, speech rhythm, emphasis, or emotion).

Previous studies have focused on the tighter inte-

[*]Work done at CNGL, School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland.

gration of ASR and MT to solve the aforementioned problems. The coupling structure proposed in (Ney, 1999) highlights the importance of information sharing between ASR and MT modules and the following studies (Mathias and Byrne, 2005; Zhou et al., 2007) evaluate using merged graphs to achieve optimal translation by integrating, searching, and combining various ASR scores and translation models. These studies can be classified as Finite State Transducers (FST)-based approaches along with the GIATI-based speech translation system (Casacuberta and Vidal, 2004; Casacuberta et al., 2004; Matusov et al., 2005a). In the FST-based approach, a tighter integration between the ASR and MT modules is accomplished by using FSTs as the basic structure to share information. The approach uses composite decoding to obtain better translation quality from source-side speech input. In this architecture source-side speech is fed into FST modules to obtain target-side translation outputs, and then the synthesis module is used to produce target-side speech.

Since the FST-based approach accomplishes a tighter integration between the ASR and MT modules, it is easier to recover from recognition errors since the ASR and MT modules are seamlessly connected. Even though a complex structure is implied on large models, the FST-based approach can be efficiently implemented with state-of-the-art MT models (Iglesias et al., 2009), which makes it applicable to run time applications.

Although speech translation has been extensively studied, most work uses words as the basic unit of information sharing. However, not much research has explored the possibility of using another unit (cf. phone) for speech translation. The word-level loses much of the rich information that is embedded in the phonetic characteristics of speech. There is substantial source-side phonetic speech information that could be used in the MT process (cf. disambiguation and recognition error recovering) but are not.

Motivated by work in (Pérez et al., 2010) that uses phonetic representations in FST for speech translation, work in (Bertoldi et al., 2008b) that utilises MT of phone-to-word translation in speech recogntion and open vocabulary speech reocogntion in (Bisani and Ney, 2005), in this paper, we introduce a phonetic representation-based approach to tackle the ASR-MT integration problem of speech translation. Instead of using words as a basic unit to carry information from ASR to MT, we utilise phones to represent source-side recognised outputs, and construct phone-to-word MT models for speech translation. Furthermore, to illustrate the potential of our approach of incorporating rich phonetic information from the source-side speech, the input phone sequences are then enriched by a phonetic confusion matrix (PCM) which is extracted from recognition outputs to represent phonetic similarities and phone-level recognition error patterns. Specifically, phone confusion networks (PCN) are included to carry information introduced by the PCM and used for the phone-to-word MT module tuning and decoding. The ultimate goal and main contribution of this work is to show the possibility and benefits of phone-level speech translation.

The remainder of this paper is organised as follows: in Section 2 we introduce the concept and architecture of speech translation of phonetic representation and the MT part named phone-to-word MT is illustrated in Section 3. Then the key module to accomplish word-to-phone conversion, namely grapheme-to-phoneme (G2P), is detailed in Section 4. In Section 5, we describe the PCM, which is used to incorporate richer source-side phonetic information. In Section 6, experiments are conducted on the IWSLT English–Chinese translation task to show the effectiveness of our proposed method, and finally conclusions and future work are then presented in Section 7.

2 ASR and MT Integration using Phonetic Representations

The overall structure of our speech translation with phonetic representation is illustrated in Figure 1. As shown in the figure, the system structure is similar to the cascaded model, but the main difference is the use of phonetic representations for the integration of ASR and MT module. This structure is looser than the FST approach as ASR and MT models are trained and tuned separately. However, since phones are used to represent ASR outputs, it provides the following merits to potentially obtain better spoken language translation outputs:

- MT uses phonetic representation of words for

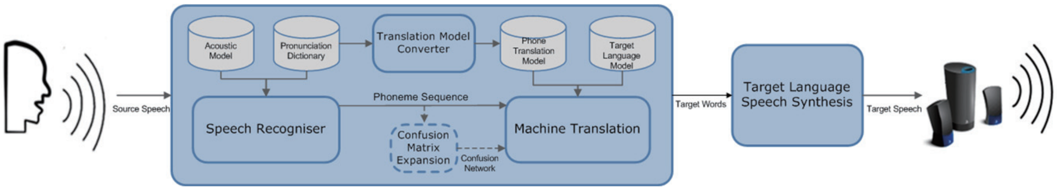

Figure 1: Phone-to-word SMT for speech-to-speech translation with phonetic representation.

translation; this contributes to homograph disambiguation (e.g. “read” has same orthographic representation for present and past but is pronounced differently).

- Even with 1-best ASR output, the MT decoder can potentially handle recognition error and OOV words in ASR language models (e.g. MT phrases help disambiguate recognition error caused by ASR).

- Domain tuning can focus on the MT side while general purpose ASR can be kept for different domains.

Both ASR and MT system are adopted in order to utilise phonetic information:

- Although the ASR module normally transcribes speech into words, state-of-the-art systems utilise phone models. Thus it is straightforward to either use a phone recogniser to convert speech into phone sequences or to convert word recognition outputs into phones. The difference here is the language model applied in the recognition process which actually affects the phone recognition rate for the subsequent MT process. For the first approach, a higher order phone language model is preferred for a better phone recognition rate (Bertoldi et al., 2008b) and will produce results that are less correlated to the language model used during recognition; while the second approach poses more constrains on the results to resemble the text that is used for language model training. In this paper, we only examine the second approach in order to compare with systems that operate on the word-level, but our method can be generally applied to any speech recogniser that outputs phone sequences.

- Most of the MT systems operates on word-level because this task is naturally related to text translation. Unlike the building process of phone-to-word MT in a monolingual task (Bertoldi et al., 2008b), parallel text corpus instead of dictionary is favoured, because state-of-the-art MT methodologies rely heavily on training from parallel data, in which word alignments play a decisive role. It is not wise to directly perform word alignments on source-side phones and target-side words since there are essential differences between them. Thus we utilise MT models trained from word representation of a corpus and convert the source-side entries into phones. Ideally this process does not change the inherit merits of MT models and is ought to produce identical results with inputs in phonetic representation. However, the decomposition from word to phones burdens the search process in MT decoding which might affect the final outputs.

Therefore, from this point of view, it is not attractive to operate the whole system on the phone-level, since theoretically our approach will underperform when compared to a system that works on the word-level. However, the merit of our approach is the flexibility to incorporate phonetic information. For example, considering the error-prone nature of ASR, we can provide multiple phone choices to phone-to-word MT if we have the information on which phones are closer to others (cf. PCM in Section 5) based on the recogniser outputs. This information is easily encoded by phones (e.g. PCN) and can be well-handled by state-of-the-art MT engines. As illustrated in Figure 1, we can embed useful phonetic information into phone sequences before they are passed into the MT module and thus the information sharing is more convenient based on the phone-level

Source entry	Target entry
life jacket	救生衣
life jacket	救生衣 就
life jacket	救生衣 就 在
life like on the	在 船上
life like on the	在 船上 的

Table 1: Source and target entries before G2P.

Source entry	Target entry
L AY F JH AE K IH T	救生衣
L AY F JH AE K IH T	救生衣 就
L AY F JH AE K IH T	救生衣 就 在
L AY F L AY K AO N DH IY	在 船上
L AY F L AY K AO N DH IY	在 船上 的

Table 2: Source and target entries after G2P.

unit between ASR and MT modules.

Note that in both the conversion of word-level ASR output into phone sequences and the building process of phone-to-word MT, the key process is the transformation from words to phones, namely G2P, which is introduced in Section 4. This module assures the validity of the whole speech translation process even when there are OOVs that cannot be handled a with source-side dictionary.

3 Phone-to-word MT

In this paper, we transform PBSMT models into phone-to-word models to accept phone sequences with the following steps:

- Perform word alignment on parallel corpus and then extract phrase table and lexical reordering table to obtain a original PBSMT model. At the same time, train target-side language model on target-side corpus for further usage.
- For each source-side entry in the phrase table and lexical reordering table, convert it into a phonetic representation with a G2P conversion module (in Section 4) to obtain a phone-to-word phrase table and lexical reordering table. Scores are left untouched. Table 1 and 2 illustrate the entries in the phrase table and lexical reordering table before and after G2P respectively. For each row in the first table, there is a corresponding row in the second table, with a different source entry after G2P conversion. For example, the word “life” and “jacket” is converted into “L AY F” and “JH AE K IH T” respectively, then the corresponding source entries are transformed from “life jacket” into “L AY F JH AE K IH T”.
- Tune the phonetic PBSMT model with phone sequences. As described in the last section, original word-level development set is transformed into source-side phone sequences with G2P module, and target side is left untouched. Then, similar to the original PBSMT building process, the phonetic PBSMT model is tuned by MERT (Och, 2003) in terms of the BLEU (Papineni et al., 2002) metric.

Note that from the previous G2P examples, since source words are transformed into their phonetic forms, the average input length is much larger than the original PBSMT (e.g. one word “jacket” is transformed into five phones “JH AE K IH T”), which implies a greater computational complexity since maximum phrase length and distortion limit ought to be increased proportionally for comparable performance. Therefore cube pruning (Chiang, 2007) for PBSMT is utilised for faster decoding.

Now we use the term “Phone Translation” (PT) to identify this phonetic MT base system and it can be enhanced with different source-side phonetic information.

4 G2P Conversion

The G2P module plays an essential role to converts words from the orthographic form (a sequence of letters) to its pronunciation representation (a sequence of phones). It is utilised to transform word-level recognition outputs into phone sequences, and to enable original PBSMT model works on source phone sequences.

Because of the irregular correspondence between spelling and pronunciation, it is a difficult task and researchers have proposed various statistical machine learning algorithms (Bisani and Ney, 2008). Following the work presented in (Zhang and Zhou, 2010), the Phrase-based Loglinear translation model is used in this paper to accomplish the G2P task with context independent modeling on the G2P mappings. The following steps are carried out to train the G2P model:

- Tranform a source-language dictionary (e.g. CMUdict[1] for English) into a parallel corpus, which contains the source-side as words and target-side as their pronunciations in phones.
- Split source-side words into separate letters for the following processes to capture the alignment properties between grapheme and phones. For example, word "jacket" is transformed into "j a c k e t" as the source-side and its corresponding target side is "JH AE K IH T" from the CMUdict. Different alignment schemes can then be learnt from these pairs (e.g. 2-to-1 alignments from "c k" to "K").
- Divide the parallel corpus into training and tuning sets. On the training set, perform word alignment, PBSMT model and language model training. Then on the tuning set, MERT is used to obtain optimal feature weights in terms of BLEU.

With the trained G2P model, input words can be transformed into phone sequences by: 1) convert the word into letters separated by spaces; 2) feed into the tuned PBSMT model for decoding outputs. The merit of this process is the ability of handing OOV words with a limited dictionary size.

5 The Use of Phone Confusion Matrix (PCM)

As depicted at the end of Section 2, the merit of the PT system is the ability to incorporate source-language phonetic information for better speech translation. In this paper, we illustrate this concept by allowing the MT module to access PCM information, which represents the source-language phonetic similarities and error-patterns of ASR. Therefore, it is possible for the MT module to restore the errors occurred in the ASR stage.

PCM has been used to represent the uncertainty brought in by the recognition (Bouselmi et al., 2005), which might because of phonetic similarity or errors in ASR. It is usually extracted by aligning the recognition outputs with transcriptions (Bertoldi et al., 2008b; Jiang and Xu, 2009). In this paper, it is extracted as follows from the word recognition outputs:

- Convert both word recognition outputs and their transcriptions into phones with the G2P module.
- Align the two corresponding phone sequences by dynamic programming, penalise on insertion, deletion and substitution.
- Calculate the confusion value of phone p with respect to phone q as in 1:

$$Conf(p,q) = \frac{M_{pq}}{N_p} \tag{1}$$

  where N_p is the total number of phone p in transcriptions, and M_{pq} is the times that phone p is aligned with phone q. Thus, given N phones, by iterating phone p and q though the entire phone set, an $N \times N$matrix is created. Note that deletion of phones is also modeled in the PCM with a special phone named $*EPS*$.
- For each phone p, the confusion values with corresponding phones are then sorted in decreasing order to obtain the PCM.

With the generated PCM, input phone sequences of MT are then enhanced into PCNs. As depicted in Figure 2, for each of the phone p_i in the phone sequence $p_1 \ldots p_{i-1} p_i p_{i+1} \ldots p_n$, the phone candidates in PCM are in a vector $[q_1, q_2, \ldots, q_m]$, then we add extra paths to allow the replacement of phones in the vector with p_i, and each edge is assigned with a weight from the confusion value calculated in equation 1. By iterating through all phones in the input, we can obtain a PCN encoded with both the original phone sequence and the PCM knowledge. To adjust the size of the generated PCNs, which are directly related to the computational complexity, the whole process is controlled by a Confusion Threshold (CT) parameter, which indicates the lowest relative confusion value obtained in equation 1 of phone candidates compared with the original phone p_i. All phone candidates below CT is pruned before feeding into the PCN construction process.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

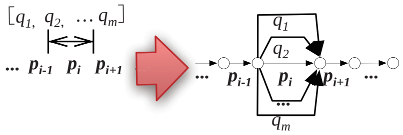

Figure 2: Convert phone sequence into PCNs.

Both the development and test set are transformed into PCNs, and the original PT system is re-tuned by MERT with a new feature comes from equation 1, which is store on the edges of PCNs. Here we call this system "Confusion Matrix Enhanced Phone Translation" (CMEPT) for comparison.

6 Experiments and Results

6.1 Experimental Setup

Experiments are carried out on the English–Chinese IWSLT[2] DIALOG task. The training corpus contains 71,725 parallel sentence pairs, and it is used to train both translation models and language models. We choose a development set with 498 sentences and a test set with 251 sentences from the available DIALOG development sets, both of which contains 7 references, while the test set comes from 1-best ASR outputs with a WER of 17.9%.

The baseline system is Moses (Koehn et al., 2007) that works on word-level, and we presented our PT and CMEPT systems that work on phone-level for comparison. The G2P module is trained on CMUdict and 101,872 entries are used for training and 3,000 entries are used for tuning. PCM is extracted from 2,060 1-best outputs and transcriptions of IWSLT English–Chinese DIALOG development sets excluding our test set. We utilised various CT parameters to construct different size of PCNs during the experiments to examine the translation quality.

For our phone-level systems, we used the maximum source entry length 57 of converted phrase table as the max-phrase-length parameter. We also used a larger distortion limit (50) for the phone-level systems and tried both stack decoding and cube pruning in the decoding stage. For comparison, we also tried different parameter settings for baseline PBSMT and report the one with the best performance .

System	CT	BLEU	TER
correct text	–	*34.86*	*43.52*
Baseline PBSMT	–	29.60	47.11
PT	–	28.43	48.18
CMEPT	0.01	30.14	46.80
	0.009	30.11	46.80
	0.008	**30.87**	**46.54**
	0.007	30.81	46.65
	0.006	30.78	46.65

Table 3: Results of PBSMT with correct text, and baseline PBSMT, PT and CMEPT systems with 1-best.

6.2 Results

The experiments are reported in terms of BLEU and TER (Snover et al., 2009). Note that the outputs using correct text for PBSMT is also showed for reference, and all other system uses 1-best ASR outputs as the input.

Table 3 compares the performance of the optimal baseline PBSMT and PT systems with different parameters, and CMEPT system with different CT values. As observed from the table, when using the IWSLT corpus, the CMEPT system with CT value 0.008 accomplishes the best performance both in terms of BLEU and TER. It outperforms the baseline by 1.27 absolute (4.29% relatively) BLEU points and 0.57 absolute (1.21% relative) TER points. It also outperforms the PT system by 2.44 absolute (8.58% relative) BLEU points and 1.64 absolute (3.40% relative) TER points. The experiment results also show that for CT values from 0.006 to 0.01, the CMEPT system consistently outperform both the baseline and PT systems. On the other hand, the PT system performs significantly lower than the baseline system by 1.17 absolute (3.95% relative) BLEU points and 1.07 absolute (2.27% relative) TER points.

6.3 Discussion and Analysis

The results showed that the basic PT system did not perform as well as the baseline PBSMT system. This is as one may expect as the search space is much larger because the model consists of phonetic representations. Furthermore, the source vocabulary contained the source language phone set (i.e. 39

[2] `http://iwslt2010.fbk.eu/`

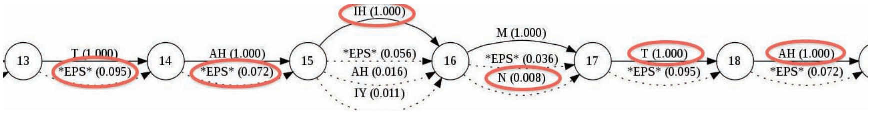

Figure 3: Example of PCN.

phone labels + 1 for deletions).

The CMEPT system was designed to incorporate additional phonetic information and its results showed that it outperformed both the PT and baseline PBSMT systems. This performance gain was a direct result of allowing MT to access the phonetic information on the source-side.

To illustrate how the CMEPT system works better than the baseline, we provide an example that contains a recognition error. For this sentence, the correct transcription is “could you please speak into the microphone” and the 1-best output is “could you please speak to him to the microphone”. The difference between “into” and “to him to” results in the translation of baseline to be “请 到 他 离 麦克 近 一点” (lit: *please go him microphone closer*), which is incorrect in this case. However, by using the constructed PCN in Figure 3, which is a subpart for “to him to”, the MT decoder choose the path with red circles to output “请 对 着 麦克风 说话” (lit: *please to microphone speak*), which is the ideal translation in this case. This result is because a subpart of the graph recovered “into” with the phones “IH N T AH”. Thus, the PCN allows the MT system to recover from ASR errors based on phonetic information, which is more straight-forward and easy to integrate than word-level systems.

7 Conclusion and future work

In this paper, we investigate the potential of using phones as the basic information sharing unit between ASR and MT for a tighter integration in speech translation. G2P modules are used to transform an original word-level PBSMT system into a phone-level system and PCM is encoded in the phone sequences as PCNs to take into account phone similarities and ASR error-patterns. The approach described was evaluated using the data from the IWSLT English–Chinese DIALOG task.

The system performed significantly better in translation quality (1.27 absolute BLEU points) than the baseline PBSMT system which indicates that phonetic information can improve speech translation systems. However, it should be noted that the BLEU score does not highlight all of the strengths of this approach, in particular:

1. The system has the potential (with the configuration of phone recogntion on the source-side) to use only one single language model, avoiding the problems associated with mismatched ASR and MT models.
2. The MT system can recover from ASR errors, as opposed to translating an ASR error into a completely incorrect word.
3. Words that cannot be translated can have their original phone sequences directly synthesised using a target language voice.

In the future, we plan to incorporate more features into the phone-to-word MT architecture and an integrated tunning approach for both ASR and MT on larger data sets. For ASR outputs, we also plan to investigate the benefits of using phoneme graphs instead of single best phone sequences. Furthermore, due to the decoding inefficiency implied in the phone input structure, it is meaningful to take more consideration with a more effective decoding algorithm for this task.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

Nicola Bertoldi, Richard Zens, Marcello Federico, and Wade Shen. 2008a. Efficient Speech Translation Through Confusion Network Decoding. In *IEEE*

Transactions on Audio, Speech, and Language Processing, 16(8), pages 1696–1705.

Nicola Bertoldi, Marcello Federico, Daniele Falavigna, Matteo Gerosa. 2008b. Fast Speech Decoding Through Phone Confusion Networks. In *Proceedings of INTERSPEECH 2008: the 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, pages 2094–2097.

Maximilian Bisani, Hermann Ney. 2008. Open vocabulary speech recognition with flat hybrid models. In *Proceedings of INTERSPEECH 2005*, Lisbon, Portugal, pages 725–728.

Maximilian Bisani, Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. In *Speech Communication*, 50(5), pages 434–451.

Ghazi Bouselmi, Dominique Fohr, Irina Illina and Jean Paul Haton. 2005. Fully Automated Non-Native Speech Recognition Using Confusion-Based Acoustic Model Integration. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, pages 345–348.

Francisco Casacuberta and Enrique Vidal. 2004. Machine Translation with Inferred Stochastic Finite-State Transducers. In *Computational Linguistics*, 30(2), pages 205–225.

Francisco Casacuberta, Hermann Ney, Franz Josef Och, Enrique Vidal, Juan Miguel Vilar, Sergio Barrachina, Ismael García-varea, David Llorens, César Martínez, Sirko Molau. 2004. Some approaches to statistical and finite-state speech-to-speech translation. In *Computer Speech & Language*, 18(1), pages 25–47.

David Chiang. 2007. Hierarchical phrase-based translation. In *Computational Linguistics*, 33(2), pages 201–228.

Gonzalo Iglesias, Adrià de Gispert, Eduardo Rodríguez Banga, William J. Byrne 2009. Hierarchical Phrase-Based Translation with Weighted Finite State Transducers. In *Proceedings of North American Chapter of the Association for Computational Linguistics - NAACL*, pages 433–441.

Jie Jiang and Bo Xu 2009. Exploring the Automatic Mispronunciation Detection of Confusable Phones for Mandarin. In *Proceedings of ICASSP: The 34th International Conference on Acoustics, Speech, and Signal Processing*, Taiwan, pages 4833–4836.

Philipp Koehn, Hieu Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, Wade Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007: demo and poster sessions*, Prague, Czech Republic, pages 177–180.

Lambert Mathias and William Byrne. 2006. Statistical Phrase-Based Speech Translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, pages 561–564.

Evgeny Matusov, S. Kanthak and Hermann Ney. 2005a. On the integration of speech recognition and statistical machine translation. In *Proceedings of the European Conference on Speech Communication and Technology*.

Evgeny Matusov, Hermann Ney and Ralph Schlüter. 2005b. Phrase-based Translation of Speech Recognizer Word Lattices Using Loglinear Model Combination. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancun, Mexico, pages 110–115.

Hermann Ney. 1999. Speech translation: Cou- pling of recognition and translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, AR, pages 517–520.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method For Automatic Evaluation of Machine Translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, pp.311-318, Philadelphia, PA.

Alicia Pérez, M. Inés Torres and Francisco Casascuberta. 2010. Potential scope of a fully-integrated architecture for speech translation. In *Proceedings of the 14th Annual Meeting of the European Association for Machine Translation*, St. Raphael, France.

Matthew Snover, Nitin Madnani, Bonnie J.Dorr and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, pages 259–268.

Ruiqiang Zhang, Genichiro Kikui, Hirofumi Yamamoto, Taro Watanabe, Frank Soong and Wai Kit Lo. 2004. A unified approach in speech-to-speech translation: Integrating features of speech recognition and machine translation. In *Proceedings of COLING 2004: The 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Rong Zhang and Bowen Zhou. 2010. Applying log linear model based context dependent machine translation techniques to grapheme-to-phoneme conversion. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, TX, pages 4634–4637.

Bowen Zhou, Laurent Besacier and Yuqing Gao. 2007. On Efficient Coupling of ASR and SMT for Speech Translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, HI, pages 101–104.