A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation

Ruiqiang Zhang and **Genichiro Kikui and Hirofumi Yamamoto**
Taro Watanabe and **Frank Soong** and **Wai Kit Lo**
ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seiika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{ruiqiang.zhang, genichiro.kikui}@atr.jp

Abstract

Based upon a statistically trained speech translation system, in this study, we try to combine distinctive features derived from the two modules: speech recognition and statistical machine translation, in a log-linear model. The translation hypotheses are then rescored and translation performance is improved. The standard translation evaluation metrics, including BLEU, NIST, multiple reference word error rate and its position independent counterpart, were optimized to solve the weights of the features in the log-linear model. The experimental results have shown significant improvement over the baseline IBM model 4 in all automatic translation evaluation metrics. The largest was for BLEU, by 7.9% absolute.

1 Introduction

Current translation systems are typically of a cascaded structure: speech recognition followed by machine translation. This structure, while explicit, lacks some joint optimality in performance since the speech recognition module and translation module are running rather independently. Moreover, the translation module of a speech translation system, a natural offspring of text-input based translation system, usually takes a single-best recognition hypothesis transcribed in text and performs standard text-based translation. Lots of supplementary information available from speech recognition, such as N-best recognition recognition hypotheses, likelihoods of acoustic and language models, is not well utilized in the translation process. The information can be effective for improving translation quality if employed properly.

The supplementary information can be exploited by a tight coupling of speech recognition and machine translation (Ney, 1999) or keeping the cascaded structure unchanged but using an integration model, log-linear model, to rescore the translation hypotheses. In this study the last approach was used due to its explicitness.

In this paper we intended to improve speech translation by exploiting these information. Moreover, a number of advanced features from the machine translation module were also added in the models. All the features from the speech recognition and machine translation module were combined by the log-linear models seamlessly.

In order to test our results broadly, we used four automatic translation evaluation metrics: BLEU, NIST, multiple word error rate and position independent word error rate, to measure the translation improvement.

In the following, in section 2 we introduce the speech translation system. In section 3, we describe the optimization algorithm used to find the weight parameters in the log-linear model. In section 4 we demonstrate the effectiveness of our technique in speech translation experiments. In the final two sections we discuss the results and present our conclusions.

2 Feature-based Log-linear Models in Speech Translation

The speech translation experimental system used in this study illustrated in Fig. 1 is a typical, statistics-based one. It consists of two major cascaded components: an automatic speech recognition (ASR) module and a statistical machine translation (SMT) module. Additionally, a third module, 'Rescore', has been added to the system and it forms a key component in the system. Features derived from ASR and SMT are combined in this module to rescore translation candidates.

Without loss of its generality, in this paper we use Japanese-to-English translation to explain the generic speech translation process. Let X denote acoustic observations of a Japanese

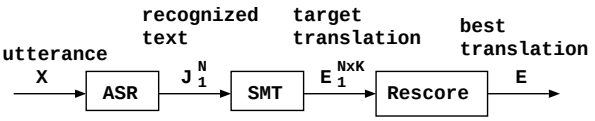

Figure 1: Current framework of speech translation

utterance, typically a sequence of short-time spectral vectors received at a frame rate of every centi-second. It is first recognized as a Japanese sentence, J. The recognized sentence is then translated into a corresponding English sentence, E.

The conversion from X to J is performed in the ASR module. Based on Bayes' rule, $P(J|X)$ can be written as

$$P(J|X) = P_{am}(X|J)P_{lm}(J)/P(X)$$

where $P_{am}(X|J)$ is the acoustic model likelihood of the observations given the recognized sentence J; $P_{lm}(J)$, the source language model probability; and $P(X)$, the probability of all acoustic observations.

In the experiment we generated a set of N-best hypotheses, $J_1^N = \{J_1, J_2, \cdots, J_N\}$ [1] and each J_i is determined by

$$J_i = \arg\max_{J \in \Omega_i} P_{am}(X|J)P_{lm}(J)$$

where Ω_i is the set of all possible source sentences excluding all higher ranked J_k's, $1 \leq k \leq i-1$.

The conversion from J to E in Fig. 1 is the machine translation process. According to the statistical machine translation formalism (Brown et al., 1993), the translation process is to search for the best sentence $\widehat{E}$ such that

$$\widehat{E} = \arg\max_{E} P(E|J) = \arg\max_{E} P(J|E)P(E)$$

where $P(J|E)$ is a translation model characterizing the correspondence between E and J; $P(E)$, the English language model probability.

In the IBM model 4, the translation model $P(J|E)$ is further decomposed into four submodels:

- Lexicon Model – $t(j|e)$: probability of a word j in the Japanese language being translated into a word e in the English language.
- Fertility model – $n(\phi|e)$: probability of a English language word e generating ϕ words.
- Distortion model – d: probability of distortion, which is decomposed into the distortion probabilities of head words and non-head words.
- NULL translation model – p_1: a fixed probability of inserting a NULL word after determining each English word.

In the above we listed seven features: two from ASR ($P_{am}(X|J)$, $P_{lm}(J)$) and five from SMT ($P(E)$, $t(j|e)$, $n(\phi|e)$, d, p_1).

The third module in Fig. 1 is to rescore translation hypotheses from SMT by using a feature-based log-linear model. All translation candidates output through the speech recognition and translation modules are re-evaluated by using all relevant features and searching for the best translation candidate of the highest score.

The log-linear model used in our speech translation process, $P(E|X)$, is

$$P_\Lambda(E|X) = \frac{exp(\sum_{i=1}^{M} \lambda_i f_i(X,E))}{\sum_{E'} exp(\sum_{i=1}^{M} \lambda_i f_i(X,E'))} \quad \Lambda = \{\lambda_1^M\} \tag{1}$$

In the Eq. 1, $f_i(X,E)$ is the logarithm value of the i-th feature; λ_i is the weight of the i-th feature. Integrating different features in the equation results in different models. In the experiments performed in section 4, four different models will be trained by increasing the number of features successively to investigate the effect of different features for improving speech translation.

In addition to the above seven features, the following features are also incorporated.

- Part-of-speech language models: English part-of-speech language models were used. POS dependence of a translated English sentence is an effective constraint in pruning English sentence candidates. In our experiments 81 part-of-speech tags and a 5-gram POS language model were used.
- Length model $P(l|E,J)$: l is the length (number of words) of a translated English sentence.
- Jump weight: Jump width for adjacent cepts in Model 4 (Marcu and Wong, 2002).
- Example matching score: The translated English sentence is matched with phrase

[1] Hereafter, J_1 is called the single-best hypothesis of speech recognition; J_1^N, the N-best hypotheses.

translation examples. A score is derived based on the count of matches (Watanabe and Sumita, 2003).

- Dynamic example matching score: Similar to the example matching score but phrases were extracted dynamically from sentence examples (Watanabe and Sumita, 2003).

Altogether, we used $M(=12)$ different features. In section 3, we review *Powell*'s algorithm (Press et al., 2000) as our tool to optimize model parameters, λ_1^M, based on different objective translation metrics.

3 Parameter Optimization Based on Translation Metrics

The denominator in Eq. 1 can be ignored since the normalization is applied equally to every hypothesis. Hence, the choice of the best translation, $\widehat{E}$, out of all possible translations, E, is independent of the denominator,

$$\widehat{E} = \arg\max_{E} \sum_{i=1}^{M} \lambda_i log P_i(X, E) \qquad (2)$$

where we write features, $f_i(X, E)$, explicitly in logarithm, $logP_i(X, E)$.

The effectiveness of the model in Eq. 2 depends upon the parameter optimization of the parameter set λ_1^M, with respect to some objectively measurable but subjectively relevant metrics.

Suppose we have L speech utterances and for each utterance, we generate N best speech recognition hypotheses. For each recognition hypothesis, K English language translation hypotheses are generated. For the l-th input speech utterance, there are then $C_l = \{E_{l_1}, \cdots, E_{l_{N\times K}}\}$ translations. All L speech utterances generate $L\times N\times K$ translations in total.

Our goal is to minimize the translation "distortion" between the reference translations, $\mathcal{R}$, and the translated sentences, $\widehat{\mathcal{E}}$.

$$\lambda_1^M = optimize \;\; \mathcal{D}(\widehat{\mathcal{E}}, \mathcal{R}) \qquad (3)$$

where $\widehat{\mathcal{E}} = \{\widehat{E}_1, \cdots, \widehat{E}_L\}$ is a set of translations of all utterances. The translation $\widehat{E}_l$ of the l-th utterance is produced by the (Eq. 2), where $E \in C_l$.

Let $\mathcal{R} = \{R_1, \cdots, R_L\}$ be the set of translation references for all utterances. Human translators paraphrased 16 reference sentences for each utterance, i.e., R_l contains 16 reference candidates for the l-th utterance.

$\mathcal{D}(\widehat{\mathcal{E}}, \mathcal{R})$ is a translation "distortion" or an objective translation assessment. The following four metrics were used specifically in this study:

- BLEU (Papineni et al., 2002): A weighted geometric mean of the n-gram matches between test and reference sentences multiplied by a brevity penalty that penalizes short translation sentences.
- NIST : An arithmetic mean of the n-gram matches between test and reference sentences multiplied by a length factor which again penalizes short translation sentences.
- mWER (Niessen et al., 2000): Multiple reference word error rate, which computes the edit distance (minimum number of insertions, deletions, and substitutions) between test and reference sentences.
- mPER: Multiple reference position independent word error rate, which computes the edit distance without considering the word order.

The BLEU score and NIST score are calculated by the tool downloadable [2].

Because the objective function in the model (Eq. 3) is not smoothed function, we used *Powell*'s search method to find a solution. The *Powell*'s algorithm used in this work is similar as the one from (Press et al., 2000) but we modified the line optimization codes, a subroutine of *Powell*'s algorithm, with reference to (Och, 2003).

Finding a global optimum is usually difficult in a high dimensional vector space. To make sure that we had found a good local optimum, we restarted the algorithm by using various initializations and used the best local optimum as the final solution.

4 Experiments

4.1 Corpus & System

The data used in this study was the Basic Travel Expression Corpus (BTEC) (Kikui et al., 2003), consisting of commonly used sentences listed in travel guidebooks and tour conversations. The corpus were designed for developing multiple language speech-to-speech translation systems. It contains four different languages: Chinese, Japanese, Korean and English. Only Japanese-English parallel data was used in this

[2] http://www.nist.gov/speech/tests/mt/

Table 1: Training, development and test data from Basic Travel Expression Corpus(BTEC)

		Japanese	English
Train	Sentences	162,318	
	Words	1,288,767	949,377
Dev.	Sentences	510	
	Words	4015	2983
Test	Sentences	508	
	Words	4112	2951

study. The speech data was recorded by multiple speakers and was used to train the acoustic models, while the text database was used for training the language and translation models.

The standard BTEC training corpus, the first file and the second file from BTEC standard test corpus #01 were used for training, development and test respectively. The statistics of corpus is shown in table 1.

The speech recognition engine used in the experiments was an HMM-based, large vocabulary continuous speech recognizer. The acoustic HMMs were triphone models with 2,100 states in total, using 25 dimensional, short-time spectrum features. In the first and second pass of decoding, a multiclass word bigram of a lexicon of 37,000 words plus 10,000 compound words was used. A word trigram was used in rescoring the results.

The machine translation system is a graph-based decoder (Ueffing et al., 2002). The first pass of the decoder generates a word-graph, a compact representation of alternative translation candidates, using a beam search based on the scores of the lexicon and language models. In the second pass an A^* search traverses the graph. The edges of the word-graph, or the phrase translation candidates, are generated by the list of word translations obtained from the inverted lexicon model. The phrase translations extracted from the Viterbi alignments of the training corpus also constitute the edges. Similarly, the edges are also created from dynamically extracted phrase translations from the bilingual sentences (Watanabe and Sumita, 2003). The decoder used the IBM Model 4 with a trigram language model and a 5-gram part-of-speech language model. The training of IBM model 4 was implemented by the GIZA++ package (Och and Ney, 2003).

4.2 Model Training

In order to quantify translation improvement by features from speech recognition and machine translation respectively, we built four log-linear models by adding features successively. The four models are:

- Standard translation model(stm): Only features from the IBM model 4 (M=5) described in section 2 were used in the log-linear models. We did not perform parameter optimization on this model. It is equivalent to setting all the λ_1^M to 1. This model was the standard model used in most statistical machine translation system. It is referred to as the baseline model.

- Optimized standard translation models (ostm): This model consists of the same features as the previous model "stm" but the parameters were optimized by *Powell*'s algorithm. We intended to exhibit the effect of parameter optimization by comparing this model with the baseline "stm".

- Optimized enhanced translation models (oetm): We incorporated additional translation features described in section 2 to enrich the model "ostm". In this model the number of the total features, M, is 10. Model parameters were optimized. We intended to show how much the enhanced features can improve translation quality.

- Optimized enhanced speech translation models (oestm): Features from speech recognition, likelihood scores of acoustic and language models, were incorporated additionally into the model "oetm". All the 12 features described in section 2 were used. Model parameters were optimized.

To optimize λ parameters of the log-linear models, we used the development data of 510 speech utterances. We adopted an N-best hypothesis approach (Och, 2003) to train λ. For each input speech utterance, $N \times K$ candidate translations were generated, where N is the number of generated recognition hypotheses and K is the number of translation hypotheses. A vector of dimension M, corresponding to multiple features used in the translation model, was generated for each translation candidate. The *Powell*'s algorithm was used to optimize these parameters. We used a large K to ensure that promising translation candidates were not

Table 2: Comparisons of single-best and N-best hypotheses of speech recognition performance in terms of word accuracy, sentence accuracy, insertion, deletion and substitution error rates

	word acc(%)	sent acc(%)	ins (%)	del (%)	sub (%)
single-best	93.5	78.7	2.0	0.8	3.6
N-best	96.1	87.0	1.2	0.3	2.2

pruned out. In the training, we set N=100 and K=1,000.

By using different objective translation evaluation metrics described in section 3, for each model we obtained four sets of optimized parameters with respect to BLEU, NIST, mWER and mPER metrics, respectively.

4.3 Translation Improvement by Additional Features

All 508 utterances in the test data were used to evaluate the models. Similar to processing the development data, the speech recognizer generated N-best (N=100) recognition hypotheses for each test speech utterance. Table 2 shows speech recognition results of the test data set in single-best and N-best hypotheses. We observed that over 8% sentence accuracy improvement was obtained from the single-best to the N-best recognition hypotheses. The recognized sentences were then translated into corresponding English sentences. 1,000 such translation candidates were produced for each recognition hypothesis. These candidates were then rescored by each of the four models with four sets of optimized parameters obtained in the training respectively. The candidates with the best score were chosen.

The best translations generated by a model were evaluated by the translation assessment metrics used to optimize the model parameters in the development. The experimental results are shown in Table 3.

In the experiments we changed the number of speech recognition hypotheses, N, to see how translation performance is changed as N. We found that the best translation was achieved when a relatively smaller set of hypotheses, N=5, was used. Hence, the values in Table 3 were obtained when N was set to 5.

We test each model by employing the single-best recognition hypothesis translations and the N-best recognition hypothesis translations.

Table 3: Translation improvement from the baseline model(stm) to the optimized enhanced speech translation model(oestm): Models are optimized using the same metric as shown in the columns. Numbers are in percentage except NIST score.

	BLEU	NIST	mWER	mPER
Single-best recognition hypothesis translation				
stm	54.2	7.5	39.8	34.8
ostm	59.0	8.9	36.2	34.0
oetm	59.2	9.9	34.3	31.5
N-best recognition hypothesis translation				
stm	55.5	7.3	39.8	35.4
ostm	61.1	8.8	36.4	33.9
oetm	61.1	10.0	34.0	31.1
oestm	62.1	10.2	33.7	29.4

The single-best translation was from the translation of the single best hypotheses of the speech recognition and the N-best hypothesis translation was from the translations of all the hypotheses produced by speech recognition.

In Table 3, we observe that a large improvement is achieved from the baseline model "stm" to the final model "oestm". The BLEU, NIST, mWER, mPER scores are improved by 7.9%, 2.7, 6.1%, 5.4% respectively. Note that a high value of BLEU and NIST score means a good translation while a worse translation for mWER and mPER. Consistent performance improvement was achieved in the single-best and N-best recognition hypotheses translations. We observed that the improvement were due to the following reasons:

- Optimization. Models with optimized parameters yielded a better translation than the models with unoptimized parameters. It can be seen by comparing the model "stm" with the model "ostm" for both the single-best and the N-best results.
- N-best recognition hypotheses. In majority of the cells in Table 3, translation performance of the N-best recognition is better than of the corresponding single-best recognition. N-best BLEU score of "ostm" improved over the single-best of "ostm" by 2.1%. However, NIST score is indifferent to the change. It appears that NIST score is insensitive to detect slight translation changes.

Table 4: Translation improvement of incorrectly recognized utterances from single-best(oetm) to N-best(oestm)

	BLEU	NIST	mWER	mPER
single-best	29.0	6.1	59.7	51.8
N-best	36.3	7.2	54.4	47.9

- Enhanced features. Translation performance is improved steadily when more features are incorporated into the log-linear models. Translation performance of model "oetm" is better than model "ostm" because more effective translation features are used. Model "oestm" is better than model "oetm" due to its enhanced speech recognition features. It confirms that our approach to integrate features from speech recognition and translation features works very well.

4.4 Recognition Improvement of Incorrectly Recognized Sentences

In previous experiments we demonstrated that speech translation performance was improved by the proposed enhanced speech translation model "oestm". In this section we want to show that this improvement is because of the significant improvement of incorrectly recognized sentences when N-best recognition hypotheses are used.

We carried out the following experiments. Only incorrectly recognized sentences were extracted for translation and re-scored by the model "oetm" for the single-best case and the model "oestm" for the N-best case. The translation results are shown in Table 4. Translation of incorrectly recognized sentences are improved significantly as shown in the table.

Because we used N-best recognition hypotheses, the log-linear model chose the recognition hypothesis among the N hypotheses which yielded the best translation. As a result, speech recognition could be improved if the higher accurate recognition hypotheses was chosen for translation. This effect can be observed clearly if we extracted the chosen recognition hypotheses of incorrectly recognized sentences. Table 5 shows the word accuracy and sentence accuracy of the recognition hypotheses selected by the translation module. The sentence accuracy of incorrectly recognized sentences was improved by 7.5%. The word accuracy was also improved.

Table 5: Recognition accuracy of incorrectly recognized utterance improved by N-best hypothesis translation.

		word acc. (%)	sent. acc. (%)
single-best		74.6	0
N-best	BLEU	76.4	7.5
	mWER	75.9	6.5

5 Discussions

As regards to integrating speech recognition with translation, a coupling structure (Ney, 1999) was proposed as a speech translation infrastructure that multiplies acoustic probabilities with translation probabilities in a one-step decoding procedure. But no experimental results have been given on whether and how this coupling structure improved speech translation.

(Casacuberta et al., 2002) used a finite-state transducer where scores from acoustic information sources and lexicon translation models were integrated together. Word pairs of source and target languages were tied in the decoding graph. However, this method was only tested for a pair of similar languages, i.e., Spanish to English. For translating between languages of different families where the syntactic structures can be quite different, like Japanese and English, rigid tying of word pair still remains to be shown its effectiveness for translation.

Our approach is rather general, easy to implement and flexible to expand. In the experiments we incorporated features from acoustic models and language models. But this framework is flexible to include more effective features. Indeed, the proposed speech translation paradigm of log-linear models have been shown effective in many applications (Beyerlein, 1998) (Vergyri, 2000) (Och, 2003).

In order to use speech recognition features, the N-best speech recognition hypotheses were needed. Using N-best could bear computing burden. However, our experiments have shown a smaller N seems to be adequate to achieve most of the translation improvement without significant increasing of computations.

6 Conclusion

In this paper we presented our approach of incorporating both speech recognition and machine translation features into a log-linear speech translation model to improve speech

translation.

Under this new approach, translation performance was significantly improved. The performance improvement was confirmed by consistent experimental results and measured by using various objective translation metrics. In particular, BLEU score was improved by 7.9% absolute.

We show that features derived from speech recognition: likelihood of acoustic and language models, helped improve speech translation. The N-best recognition hypotheses are better than the single-best ones when they are used in translation. We also show that N-best recognition hypothesis translation can improve speech recognition accuracy of incorrectly recognized sentences.

The success of the experiments owes to the use of statistical machine translation and log-linear models so that various of effective features can be jointed and balanced to output the optimal translation results.

Acknowledgments

We would like to thank for assistance from Eiichiro Sumita, Yoshinori Sagisaka, Seiichi Yamamoto and the anonymous reviewers.

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology of Japan entitled "A study of speech dialogue translation technology based on a large corpus".

References

Peter Beyerlein. 1998. Discriminative model combination. In *Proc.of ICASSP'1998*, volume 1, pages 481–484.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Francisco Casacuberta, Enrique Vidal, and Juan M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. In *Proc. of speech-to-speech translation workshop*, pages 39–44, Philadelphia, PA, July.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc.of EUROSPEECH'2003*, pages 381–384, Geneva.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP-2002*, Philadelphia, PA, July.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *Proc. of ICASSP'1999*, volume 1, pages 517–520, Phoenix, AR, March.

Sonja Niessen, Franz J. Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proc.of the LREC (2000)*, pages 39–45, Athens, Greece, May.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL'2003*, pages 160–167.

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

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2000. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

Nicola Ueffing, Franz Josef Och, and Hermann Ney. 2002. Generation of word graphs in statistical machine translation. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP02)*, pages 156–163, Philadelphia, PA, July.

Dimitra Vergyri. 2000. Use of word level side information to improve speech recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000.*

Taro Watanabe and Eiichiro Sumita. 2003. Example-based decoding for statistical machine translation. In *Machine Translation Summit IX*, pages 410–417, New Orleans, Louisiana.