# NICT-ATR Speech-to-Speech Translation System

**Eiichiro Sumita** **Tohru Shimizu** **Satoshi Nakamura**

National Institute of Information and Communications Technology
&
ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
eiichiro.sumita, tohru.shimizu & satoshi.nakamura@atr.jp

## Abstract

This paper describes the latest version of speech-to-speech translation systems developed by the team of NICT-ATR for over twenty years. The system is now ready to be deployed for the travel domain. A new noise-suppression technique notably improves speech recognition performance. Corpus-based approaches of recognition, translation, and synthesis enable coverage of a wide variety of topics and portability to other languages.

## 1 Introduction

Speech recognition, speech synthesis, and machine translation research started about half a century ago. They have developed independently for a long time until speech-to-speech translation research was proposed in the 1980's. The feasibility of speech-to-speech translation was the focus of research at the beginning because each component was difficult to build and their integration seemed more difficult. After groundbreaking work for two decades, corpus-based speech and language processing technology have recently enabled the achievement of speech-to-speech translation that is usable in the real world.

This paper introduces (at ACL 2007) the state-of-the-art speech-to-speech translation system developed by NICT-ATR, Japan.

## 2 SPEECH-TO-SPEECH TRANSLATION SYSTEM

A speech-to-speech translation system is very large and complex. In this paper, we prefer to describe recent progress. Detailed information can be found in [1, 2, 3] and their references.

### 2.1 Speech recognition

To obtain a compact, accurate model from corpora with a limited size, we use MDL-SSS [4] and composite multi-class N-gram models [5] for acoustic and language modeling, respectively. MDL-SSS is an algorithm that automatically determines the appropriate number of parameters according to the size of the training data based on the Maximum Description Length (MDL) criterion. Japanese, English, and Chinese acoustic models were trained using the data from 4,200, 532, and 536 speakers, respectively. Furthermore, these models were adapted to several accents, e.g., US (the United States), AUS (Australia), and BRT (Britain) for English. A statistical language model was trained by using large-scale corpora (852 k sentences of Japanese, 710 k sentences of English, 510 k sentences of Chinese) drawn from the travel domain.

Robust speech recognition technology in noisy situations is an important issue for speech translation in real-world environments. An MMSE (Minimum mean square error) estimator for log Mel-spectral energy coefficients using a GMM (Gaussian Mixture Model) [6] is introduced for suppressing interference and noise and for attenuating reverberation.

Even when the acoustic and language models are trained well, environmental conditions such as variability of speakers, mismatches between the training and testing channels, and interference from environmental noise may cause recognition errors. These utterance recognition errors can be rejected by tagging them with a low confidence value. To do this we introduce generalized word



*Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 25–28,
Prague, June 2007. 

posterior probability (GWPP)-based recognition error rejection for the post processing of the speech recognition [7, 8].

### 2.2 Machine translation

The translation modules are automatically constructed from large-scale corpora: (1) TATR, a phrase-based SMT module and (2) EM, a simple memory-based translation module. EM matches a given source sentence against the source language parts of translation examples. If an exact match is achieved, the corresponding target language sentence will be output. Otherwise, TATR is called up. In TATR, which is built within the framework of feature-based exponential models, we used the following five features: phrase translation probability from source to target; inverse phrase translation probability; lexical weighting probability from source to target; inverse lexical weighting probability; and phrase penalty.

Here, we touch on two approaches of TATR: novel word segmentation for Chinese, and language model adaptation.

We used a subword-based approach for word segmentation of Chinese [9]. This word segmentation is composed of three steps. The first is a dictionary-based step, similar to the word segmentation provided by LDC. The second is a subword-based IOB tagging step implemented by a CRF tagging model. The subword-based IOB tagging achieves a better segmentation than character-based IOB tagging. The third step is confidence-dependent disambiguation to combine the previous two results. The subword-based segmentation was evaluated with two different data from the Sighan Bakeoff and the NIST machine translation evaluation workshop. With the data of the second Sighan Bakeoff[1], our segmentation gave a higher F-score than the best published results. We also evaluated this segmentation in a translation scenario using the data of NIST translation evaluation[2] 2005, where its BLEU score[3] was 1.1% higher than that using the LDC-provided word segmentation.

The language model that is used plays an important role in SMT. The effectiveness of the language model is significant if the test data happen to have the same characteristics as those of the training data for the language models. However, this coincidence is rare in practice. To avoid this performance reduction, a topic adaptation technique is often used. We applied this adaptation technique to machine translation. For this purpose, a "topic" is defined as clusters of bilingual sentence pairs. In the decoding, for a source input sentence, f, a topic T is determined by maximizing $P(f|T)$. To maximize $P(f|T)$ we select cluster T that gives the highest probability for a given translation source sentence f. After the topic is found, a topic-dependent language model $P(e|T)$ is used instead of $P(e)$, the topic-independent language model. The topic-dependent language models were tested using IWSLT06 data[4]. Our approach improved the BLEU score between 1.1% and 1.4%. The paper of [10] presents a detailed description of this work.

### 2.3 Speech synthesis

An ATR speech synthesis engine called XIMERA was developed using large corpora (a 110-hour corpus of a Japanese male, a 60-hour corpus of a Japanese female, and a 20-hour corpus of a Chinese female). This corpus-based approach makes it possible to preserve the naturalness and personality of the speech without introducing signal processing to the speech segment [11]. XIMERA's HMM (Hidden Markov Model)-based statistical prosody model is automatically trained, so it can generate a highly natural F0 pattern [12]. In addition, the cost function for segment selection has been optimized based on perceptual experiments, thereby improving the naturalness of the selected segments [13].

## 3 EVALUATION

### 3.1 Speech and language corpora

We have collected three kinds of speech and language corpora: BTEC (Basic Travel Expression Corpus), MAD (Machine Aided Dialog), and FED (Field Experiment Data) [14, 15, 16, and 17]. The BTEC Corpus includes parallel sentences in two languages composed of the kind of sentences one might find in a travel phrasebook. MAD is a dialog corpus collected using a speech-to-speech translation system. While the size of this corpus is relatively limited, the corpus is used for adaptation and

[1] http://sighan.cs.uchicago.edu/bakeoff2005/
[2] http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html
[3] http://www.nist.gov/speech/tests/mt/resources/scoring.htm
[4] http://www.slt.atr.jp/IWSLT2006/

evaluation. FED is a corpus collected in Kansai International Airport uttered by travelers using the airport.

### 3.2 Speech recognition system

The size of the vocabulary was about 35 k in canonical form and 50 k with pronunciation variations. Recognition results are shown in Table 1 for Japanese, English, and Chinese with a real-time factor[5] of 5. Although the speech recognition performance for dialog speech is worse than that for read speech, the utterance correctness excluding erroneous recognition output using GWPP [8] was greater than 83% in all cases.

| | | BTEC | MAD | FED |
|---|---|---|---|---|
| *Characteristics* | | Read speech | Dialog speech (Office) | Dialog speech (Airport) |
| *# of speakers* | | 20 | 12 | 6 |
| *# of utterances* | | 510 | 502 | 155 |
| *# of word tokens* | | 4,035 | 5,682 | 1,108 |
| *Average length* | | 7.9 | 11.3 | 7.1 |
| *Perplexity* | | 18.9 | 23.2 | 36.2 |
| *Word accuracy* | *Japanese* | 94.9 | 92.9 | 91.0 |
| | *English* | 92.3 | 90.5 | 81.0 |
| | *Chinese* | 90.7 | 78.3 | 76.5 |
| *Utterance correctness* | *All* | 82.4 | 62.2 | 69.0 |
| | *Not rejected* | 87.1 | 83.9 | 91.4 |

**Table 1 Evaluation of speech recognition**

### 3.3 Machine Translation

The mechanical evaluation is shown, where there are sixteen reference translations. The performance is very high except for English-to-Chinese (Table 2).

| | BLEU |
|---|---|
| *Japanese-to-English* | 0.6998 |
| *English-to-Japanese* | 0.7496 |
| *Japanese-to-Chinese* | 0.6584 |
| *Chinese-to-Japanese* | 0.7400 |
| *English-to-Chinese* | 0.5520 |
| *Chinese-to-English* | 0.6581 |

**Table 2 Mechanical evaluation of translation**

The translation outputs were ranked A (perfect), B (good), C (fair), or D (nonsense) by professional translators. The percentage of ranks is shown in Table 3. This is in accordance with the above BLEU score.

| | A | AB | ABC |
|---|---|---|---|
| *Japanese-to-English* | 78.4 | 86.3 | 92.2 |
| *English-to-Japanese* | 74.3 | 85.7 | 93.9 |
| *Japanese-to-Chinese* | 68.0 | 78.0 | 88.8 |
| *Chinese-to-Japanese* | 68.6 | 80.4 | 89.0 |
| *English-to-Chinese* | 52.5 | 67.1 | 79.4 |
| *Chinese-to-English* | 68.0 | 77.3 | 86.3 |

**Table 3 Human Evaluation of translation**

## 4 System presented at ACL 2007

The system works well in a noisy environment and translation can be performed for any combination of Japanese, English, and Chinese languages. The display of the current speech-to-speech translation system is shown below.

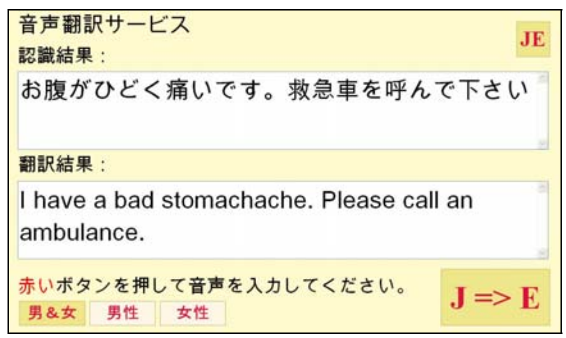


**Figure 1 Japanese-to-English Display of NICT-ATR Speech-to-Speech Translation System**

## 5 CONCLUSION

This paper presented a speech-to-speech translation system that has been developed by NICT-ATR for two decades. Various techniques, such as noise suppression and corpus-based modeling for both speech processing and machine translation achieve robustness and portability.

The evaluation has demonstrated that our system is both effective and useful in a real-world environment.

[5] The real time factor is the ratio to an utterance time.

## References


[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. The ATR multilingual speech-to-speech translation system. IEEE Trans. on Audio, Speech, and Language Processing, 14, No. 2:365–376, 2006.

[2] T. Shimizu, Y. Ashikari, E. Sumita, H. Kashioka, and S. Nakamura, “Development of client-server speech translation system on a multi-lingual speech communication platform,” Proc. of the International Workshop on Spoken Language Translation, pp. 213-216, Kyoto, Japan, 2006.

[3] R. Zhang, H. Yamamoto, M. Paul, H. Okuma, K. Yasuda, Y. Lepage, E. Denoual, D. Mochihashi, A. Finch, and E. Sumita, “The NiCT-ATR Statistical Machine Translation System for the IWSLT 2006 Evaluation,” Proc. of the International Workshop on Spoken Language Translation, pp. 83-90, Kyoto, Japan , 2006.

[4] T. Jitsuhiro, T. Matsui, and S. Nakamura. Automatic generation of non-uniform context-dependent HMM topologies based on the MDL criterion. In Proc. of Eurospeech, pages 2721–2724, 2003.

[5] H. Yamamoto, S. Isogai, and Y. Sagisaka. Multi-class composite N-gram language model. Speech Communication, 41:369–379, 2003.

[6] M. Fujimoto and Y. Ariki. Combination of temporal domain SVD based speech enhancement and GMM based speech estimation for ASR in noise - evaluation on the AURORA II database and tasks. In Proc. of Eurospeech, pages 1781–1784, 2003.

[7] F. K. Soong, W. K. Lo, and S. Nakamura. Optimal acoustic and language model weight for minimizing word verification errors. In Proc. of ICSLP, pages 441–444, 2004

[8] W. K. Lo and F. K. Soong. Generalized posterior probability for minimum error verification of recognized sentences. In Proc. of ICASSP, pages 85–88, 2005.

[9] R. Zhang, G. Kikui, and E. Sumita, “Subword-based tagging by conditional random fields for Chinese word segmentation,” in Companion volume to the proceedings of the North American chapter of the Association for Computational Linguistics (NAACL), 2006, pp. 193–196.

[10] H. Yamamoto and E. Sumita, “Online language model task adaptation for statistical machine translation (in Japanese),” in FIT2006, Fukuoka, Japan, 2006, pp. 131–134.

[11] H. Kawai, T. Toda, J. Ni, and M. Tsuzaki. XIMERA: A new TTS from ATR based on corpus-based technologies. In Proc. of 5th ISCA Speech Synthesis Workshop, 2004.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In Proc. of ICASSP, pages 1215–1218, 2000.

[13] T. Toda, H. Kawai, and M. Tsuzaki. Optimizing sub-cost functions for segment selection based on perceptual evaluation in concatenative speech synthesis. In Proc. of ICASSP, pages 657–660, 2004.

[14] T. Takezawa and G. Kikui. Collecting machine –translation-aided bilingual dialogs for corpus-based speech translation. In Proc. of Eurospeech, pages 2757–2760, 2003.

[15] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In Proc. Of Eurospeech, pages 381–384, 2003.

[16] T. Takezawa and G. Kikui. A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation. In Proc. of LREC, pages 1589–1592, 2004.

[17] G. Kikui, T. Takezawa, M. Mizushima, S. Yamamoto, Y. Sasaki, H. Kawai, and S. Nakamura. Monitor experiments of ATR speech-to-speech translation system. In Proc. of Autumn Meeting of the Acoustical Society of Japan, pages 1–7–10, 2005, in Japanese.