# 2U-03 Speech Translation System for Travel Conversation

*Akitoshi Okumura, Kiyoshi Yamabana, Shin-ichi Doi, Shinsuke Sakai, Shin-ichiro Kamei,*
*Kazuhiro Takahashi, Ken Hanazawa, Kenji Satoh, Takao Watanabe*
NEC, Computer & Communication Media Research

## 1. Introduction

In order to develop a practical speech translation system which can help oral communication between speakers of different languages during their travel abroad, the problem of the limitation in the variety of acceptable situations and expressions needs to be solved. For a relatively small-scale domain, an approach which uses domain-specific concept(inter-lingua) such as 'speech act' or 'semantic frame' is effective[1]. However, this approach would not be applicable to a larger scale of domain because the design of such domain-specific concept which is rather difficult.

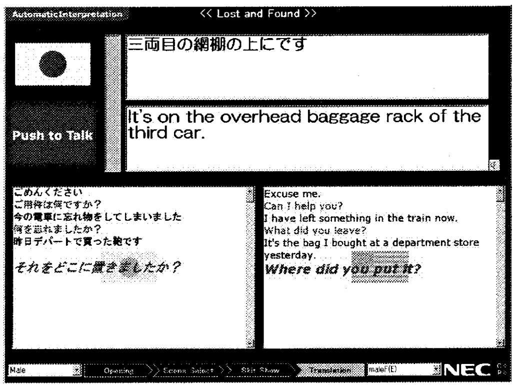
On the other hand, speech recognition technology for dictating text and machine translation technology for document translation are widely available today, which are basically designed for general purpose, that is, for domain–independent use. One approach to deal with a large-scale domain is to combine these technologies through the interface expressed as a text form, where a speech recognition system and a translation system deal with conversational speech and domain specific expressions.

One idea for this is to derive the knowledge for the speech recognition system and the translation system from a large amount of domain knowledge base, in particular "domain corpus". This idea should be hopeful if a sufficiently large amount of corpus is collected. However the amount of the corpus which can be collected is inevitably insufficient for a large-scale domain.

The adopted approach utilizes the general linguistic knowledge as well as the domain-specific linguistic knowledge. In speech recognition, on the basis of statistical language modeling, a language model was developed using both the domain knowledge, that is domain corpora, and general linguistic knowledge, and incorporate it into large vocabulary continuous speech recognition. In translation, a newly developed lexicalized grammar formalism is suitable to handle both pattern-like expressions specific to the domain conversations as well as general expressions. Direct language-pair based translation approach was adopted instead of inter-lingua approach.

## 2. System Overview

The system assists travelers with communication in a wide variety of situations. Users can have their speech simultaneously translated in real-time by a mobile PC from either Japanese or English to the other. Together with a 50,000 Japanese and 10,000 English word vocabulary, the software allows users to speak naturally without restriction, as shown in Figure 1.



三両目の網棚の上にです

It's on the overhead baggage rack of the third car.

ごめんください
ご用件は何ですか？
今の電車に忘れ物をしてしまいました
何を忘れましたか？
昨日デパートで買った鞄です
*それをどこに置きましたか？*

Excuse me.
Can I help you?
I have left something in the train now.
What did you leave?
It's the bag I bought at a department store yesterday.
***Where did you put it?***

Figure. 1 Example of a display

To reduce misunderstanding in the conversation between users talking with each other through the system and to avoid the halt in the conversation, the system accepts input of any utterance unit besides a sentence, namely a fragment of a sentence, a phrase, or a word. For each utterance, translated result is obtained in real-time. Users can make sure the recognized result is correct. Users can also request the system to re-translate after editing the recognized results utilizing the functions such as deleting a portion of the recognized result or inserting text by an additional utterance.

The system consists of four modules: speech recognition, translation, speech synthesis and system. Recognized or translated results are passed between modules in the form of a text with the supplemental information such as pausing. Japanese speech synthesis module uses a text-to-speech conversion software developed in NEC. English speech synthesis module uses a commercially available software.

System requirements for the software include a mobile PC with a Pentium II-class processor (400MHz) running either Windows 98 or Windows NT and 128MB of RAM.

Speech Translation System for Travel Conversation
Akitoshi Okumura, Kiyoshi Yamabana, Shin-ichi Doi, Shinsuke Sakai, Shin-ichiro Kamei,Kazuhiro Takahashi, Ken Hanazawa, Kenji Satoh, Takao Watanabe
a-okumura@bx.jp.nec.com
NEC Corporation,4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216-8555, Japan

## 3. Speech Recognition

Speech recognition module performs speaker-independent large-vocabulary continuous speech recognition of conversational Japanese and English. The module consists of an acoustic model, a language model, a word dictionary and a search engine. The acoustic model used was designed domain independently. The language model was designed for travel conversation. The language model contains a bigram language model and a trigram language model. The search engine performs two-stage processing. On the first stage, Viterbi beam search is performed to decode input speech to generate a word candidate graph using the acoustic model and the bigram language model. On the second stage, the engine performs a search to find the optimal word sequence using the trigram language model.

For acoustic modeling, triphone-context phone HMM was adopted. Speaker-independent recognition was made possible by training the model with a large speech corpus. The recognition module also has a speaker adaptation capability. It is possible to adapt the acoustic models efficiently to the speaker just using as few as five utterances.

The speech recognition module was evaluated using native speakers' clean utterances in speaker-independent mode (gender-dependent). Test set perplexity was 26.8 for English and 33.0 for Japanese. In English evaluation, 88.9 % of the words were correctly recognized for 20 speakers where each uttered 190 sentences (6 word length on average). The word accuracy after penalizing insertion errors was 85.4%. In Japanese evaluation, 96.2% of the words were correctly recognized for 20 speakers where each uttered 200 sentences (6.8 word length on average). The word accuracy was 95.8%.

## 4. Translation

In translation of conversations, a translation module is required to cope with highly word-specific phenomena, including various colloquial and idiomatic expressions. Handling idiosyncratic word behavior is also important to improve the translation quality for the target domain. In addition, translation module is required to cover a wide range of input sentences.

To achieve both broad coverage for general input and high quality for the target domain, we employed a rule-based method that allows writing of both general abstract rules and example-like concrete patterns in a unified framework. Precisely we adopted a strong lexicalization approach to the grammar [3] where all grammar rules(trees) are associated with at least one word, making all the rules lexical rules.

The new strongly lexicalized grammar formalism, known as Lexicalized Tree AutoMata-based Grammar (LTAMG)[2], lexicalizes (part of) tree operations as well as the trees themselves. In this formalism, each word has a tree automaton (tree accepter) that describes how to combine the elementary trees to get the whole set of trees associated with that word. This lexicalized tree automata (LTA) allow powerful and flexible control over the tree growth. Even the complex pattern-like trees having variable part inside can be easily described by the LTA without considering side effects to other words. Another advantage of the method is use of a simple chart-parsing algorithm which is a straightforward extension of the context-free grammar case.

A bilingual evaluators classified the results of 500 travel conversation sentences into four categories: "Natural" (accurate translation), "Good" (there is no syntactic error and the meaning is conveyed without error), "Understandable" (loose translation but still the core meaning can be understood) and "Bad" (corrupt sentence or has an error that causes misunderstanding). The average length of the input sentences was 6.2 words for the English sentences, and 8.9 words (morphemes) for the Japanese sentences. A sentence understanding rate, that is, the ratio of the sentences other than "Bad", of 83 – 87 % was obtained.

## 5. Conclusion

The developed speech translation system runs on a mobile PC and helps oral communication between Japanese and English speakers in various situations during their travel abroad. In order to allow a wide range of expressions and topics in the application domain, the system utilizes the general linguistic knowledge as well as the domain-specific linguistic knowledge. Speech recognition module performs speaker-independent large-vocabulary (50,000 Japanese words and 10,000 English words) continuous speech recognition. Translation module performs syntax directed translation based on a new lexicalized grammar rule formalism.

For the preliminary evaluation for relative short and clean sentences, a word accuracy of 85 – 96% for speech recognition and a sentence understanding rate of 83 – 87% for translation were obtained as a result. We expect that the performance will be further improved by expanding the grammar with regard to domain-specific or colloquial expressions, which were not yet described in the grammar and caused the incorrect translation. We will also evaluate the usability of the system as aids for cross lingual communication.

## 6. Rreferences

[1]Watanabe, T et al. An experimental automatic interpretation system: INTERTALKER, Proc. Acouts. Soc. Japan, Spring Meeting, pp.101-102 (1992) (In Japanese)

[2]Schabes, Y. et al. Parsing Strategies with `Lexicalized' Grammars. COLING'88, pp.578-583 (1988).

[3]Yamabana, K. et al. Lexicalized Tree Automata-based Grammars for Translating Conversational Texts, To appear COLING 2000 (2000)