Speech-to-Speech Translation Software on PDAs for Travel Conversation

By Ryosuke ISOTANI,* Kiyoshi YAMABANA,* Shinichi ANDO,† Ken HANAZAWA,* Shin-ya ISHIKAWA* and Ken-ichi ISO*

ABSTRACT We present an automatic speech-to-speech translation system for Personal Digital Assistants (PDAs) that helps oral communication between Japanese and English speakers in various situations while traveling. Our own compact large-vocabulary continuous speech recognition engine and compact translation engine based on a lexicalized grammar provided the basis for the Japanese/English bi-directional speech translation system that works with limited computational resources.

KEYWORDS Speech-to-speech translation, PDA (Personal Digital Assistant), Speech recognition, Machine translation, Large vocabulary

1. INTRODUCTION

Automatic speech translation is one of the most promising applications of the speech and language technology. In order to develop a practical speech translation system for users while traveling outside the country, the limitation in a variety of acceptable situations and expressions needs to be solved. Our aim was to develop a system that allows a wide range of topics and expressions and concurrently provides high-quality translation. We created a speech translation software meeting these requirements that runs on notebook PCs[1,2].

Compact implementation of the software is another issue of concern, since it aims to assist travelers. Our previous system for notebook PCs was designed in consideration of portability, but a version which worked on smaller devices was desirable. Recently, Personal Digital Assistants (PDAs) are becoming popular. Porting speech translation system on PDAs will make it more usable. Speech translation systems on PDAs already existed but they were limited to pre-registered phrases, and speech input was used in search function only[3].

In this paper, we present a speech-to-speech translation system for PDAs that possesses a large vocabulary (several ten thousands of words) and can accept a wide variety of conversational expressions in a travel situation (Fig. 1). In order to realize such a system, we made the modules of our previous system more compact, and some of them were re-built from scratch. In the following sections we give an overview of the system and the details of each module, focusing mainly on speech recognition and translation modules.

2. SYSTEM OVERVIEW

The system performs bi-directional speech translation between Japanese and English in a travel situation. The user first chooses his/her language and speaks to the system. The recognized text is displayed and translated to the other language. The user can edit the recognized text before translation. The translated text is displayed and spoken in a synthesized voice. Examples of phrases managed by the system include “Where is the station?,” “My bag was stolen downtown,” “Your boarding gate has been changed to gate seven,” and so on.

*Multimedia Research Laboratories
†Internet Solution Platform Development Division

Fig. 1 Speech translation system on a PDA.
The system is implemented as application software on PDAs running Pocket PC 2002. The requirements for the system include StrongARM* 206MHz for the CPU, 64MB of RAM and a memory card as read-only storage. No other external devices are needed.

As shown in Fig. 2, the system consists of speech recognition, translation and speech synthesis modules. A system integration module controls these modules. All the modules have been developed on our proprietary technology except the English speech synthesis, for which we used a commercially available module.

3. SPEECH RECOGNITION MODULE

The speech recognition module performs large-vocabulary continuous speech recognition of conversational Japanese and English based on HMMs and a statistical language model. The recognition result is passed to the translation module. The reading and duration of each word are added to the result to be used for disambiguation in the translation process.

As shown in Fig. 3, the speech recognition module consists of an acoustic model, a language model, a word dictionary, and a search engine. The search engine is composed of two passes: word graph generation and optimal word sequence search. The search engine is language-independent while the acoustic model, the language model, and the word dictionary are language-dependent.

To implement on a PDA, efforts are focused on downsizing of the models and reducing the computational cost and work memory of the search engine. Downsizing of the models contributes to reduction of computational cost as well as memory usage.

3.1 Acoustic Model

A speech signal is sampled at 11kHz, with MFCC analysis frame rate of 11ms. Spectral subtraction (SS) and cepstrum mean normalization (CMN) are applied to remove stationary additive and multiplicative noises. The feature set including MFCC and energy with their time derivatives is transformed by linear discriminant analysis (LDA).

The speech recognizer supports triphone HMMs with tree-based state clustering on phonetic contexts. The state emission probability is represented by gaussian mixtures with diagonal covariance matrices.

---

*StrongARM is a trademark of ARM Ltd.
covering the subsidiary leaf node gaussians. For an input feature vector, the likelihood calculation is made only for the gaussians whose likelihoods are assumed to be large by traversing the tree from the root node. The method increased the computation speed by more than ten times with the accuracy loss kept to minimum.

3.2 Language Model

To cover the wide variety of colloquial and domain-specific expressions in travel conversation, we developed large text corpora of travel conversation in Japanese and English and trained word n-gram models using them. The corpora contained a text corpus of travel conversation in various situations such as hotels, restaurants, shopping, transportation, entertainment, and so on. The corpora also contained a general conversational expression corpus comprising expressions specific to oral communication. The total size of the corpora is about one hundred thousand sentences for each language.

Word bigram and word trigram models were trained using these corpora. Class n-grams were used for smoothing word n-grams. The classes were defined on the basis of parts of speech and partly on manually defined semantic classes specific to the travel conversation domain.

To reduce memory usage of the language model parameters, each probability value was quantized and stored in one byte. A preliminary experiment showed that this quantization causes no performance degradation.

The recognition dictionary was made up with the words in the training corpora, supplemented by the words appearing in high frequency in general text. The vocabulary size is 50,000 words for Japanese and 20,000 words for English.

3.3 Search Engine (Decoder)

The search engine performs two-stage processing. In the first stage, input speech is decoded to generate a word candidate graph using the acoustic model and the bigram language model. In the second stage, the graph is searched to find the optimal word sequence using the trigram language model.

The first stage performs a frame-synchronous Viterbi beam search on a lexical prefix tree. A node of the lexical tree is a phoneme and is expressed in one byte. Additional information is stored only when a node is a branch or leaf node, thus allowing a compact representation of the lexicon. Each phoneme node is expanded to triphones on demand during the search process. The cross-word context dependency can be handled. Garbage collection is carried out periodically on the partial word graph generated during search to suppress increase of work memory.

The memory consumption of the speech recognition module is about 14MB in total of both languages after the startup. The work memory is around 1MB.

4. TRANSLATION MODULE

The translation module accepts a recognized expression from the speech recognition modules, and performs bi-directional translation between Japanese and English. The output is sent to the corresponding speech synthesis module and converted into speech.

In travel conversation, colloquial expressions are frequently used, which seldom appear in written texts such as newspapers. The translation module has to deal with highly word-specific phenomena included in such colloquial and idiomatic expressions, as well as general expressions obeying more abstract, standard grammar. In other words, the module is required to deal with both instance-specific example-like knowledge and abstract rule-like knowledge at the same time.

To satisfy this requirement, we have developed and implemented the Lexicalized Tree AutoMata-based Grammars (LTAM Grammars), a lexicalized grammar formalism, as a framework for grammar writing. This method is in line with the strongly lexicalization approach to grammar, where each grammar rule (tree) is associated with at least one word, making all the rules lexical. An advantage of the LTAM Grammars to other strongly lexicalized grammars is the existence of a simple bottom-up chart-parsing algorithm, which is a natural extension of the context-free grammar case.

Figure 4 shows the bi-directional translation module we have developed. The system uses a combined lexicalized grammar dictionary, instead of having the grammar and the dictionary separately. Large text corpora of travel conversation are used to build the bilingual dictionary and to improve the translation quality.

Since the module performs a breadth-first search for the syntactic analysis, efficient reduction/sharing of the syntactic ambiguities has been an important issue for the compact implementation.

4.1 LTAM Grammars

The basic idea of the LTAM Grammar is to hold a set of constituent trees and an automaton in a word to express the set of trees headed by the word. Each word has a set of constituent trees where one of the
leaf nodes is marked as self, and an automaton that accepts sequences of the constituent trees. A constituent-tree sequence is identified with a tree built by concatenating the trees in a bottom-up manner, identifying a root node and a self leaf node according to the order in the sequence.

Figure 5 shows a dictionary example for the verb "eat". There are three constituent trees $T_1$, $T_2$, and $T_3$ associated with the word "eat" as shown in (a). The tree automaton is a finite-state automaton as shown in (b). It accepts a sequence of trees $T_1T_2*T_3$, which is identified with the tree set in Fig. 6, including a countable number of syntax trees.

Each word has its own tree automaton, but a static sharing mechanism of the tree automata and the constituent trees is provided to reduce the cost and the size of the grammar.

4.2 Bilingual Grammars and Dictionaries for Travel Domain

The module uses reinforced bilingual lexical knowledge (Lexicalized grammar dictionary and related resources), which has been developed for our former speech translation system. The Japanese to English translation dictionary contains about a hundred and fifty thousand words, and the English to Japanese dictionary contains about seventy thousand words.

The Japanese to English translation is based on general bilingual grammar rules, and further developed for conversational expressions as will appear in a dialogue (e.g. ellipsis, idioms, fixed expressions, speech act idioms, compound words). The English to Japanese translation grammar dictionary is developed with general grammar rules and aimed to achieve high quality in the travel domain by covering a wide variety of conversational expressions, such as various forms of colloquial idioms, imperatives, requests and polite expressions.

4.3 Compact Translation Engine

The compact translation engine performs bidirectional translation between Japanese and English. The translation direction is determined by the grammar-dictionary it uses.

The translation module first performs morphological analysis to build an initial word lattice. Then for each edge in the lattice, the feature structure and the tree automata are loaded from the dictionary. The parser performs left-to-right bottom-up chart parsing in a breadth-first manner. During parsing, each edge is assigned a score calculated from the lexical, syntactic, and semantic information. An edge with the highest-score is chosen and then the generation steps proceed in a top-down manner. After the morphological generation, the translation result is passed to the speech synthesis module.
In a strongly lexicalized grammar, the grammatical knowledge is localized in the word dictionary, which is an advantage for compact implementation[8]. In our framework, static tree/tree-automata sharing and dynamic calculation of Shared Packed Forest also efficiently reduce the resource requirement. Implementation of a beam-search and a lazy-copy mechanism of the feature structures also contribute to efficient use of the CPU and the memory.

The memory consumption of the translation module (both the English-Japanese and Japanese-English) is about 8MB after the startup of the module. The work memory is about 1 to 4MB, depending on the input data.

5. SYSTEM INTEGRATION MODULE

The system integration module controls all the engine modules providing coordinated functionality as a speech translator. It also establishes the interface between the engine modules. Some of the engine modules are implemented as separate processes, which communicate with the system integration module through shared memory-mapping objects.

From the speech recognition module to the translation module, information on word segmentation, reading, and duration is passed on in addition to the text. This information is used for disambiguation in the later module, such as homograph disambiguation. Non-textual information is also passed from the translation module to the Japanese speech synthesis module for better speech quality.

In order to facilitate a quick switch of the translation direction, all the engines are kept up and ready once the translator is started. Since the operating system (Pocket PC 2002) limits the process address space to 32MB with various restrictions on the memory map, the module has been designed and implemented with great care concerning the space allocation policy.

6. DISCUSSION

Evaluation experiments of the speech recognition module were conducted using 1,800 utterances by ten male speakers for Japanese and English, respectively. The word accuracy was 95% for Japanese and 87% for English, which is roughly the same as our former PC-based speech translation systems[1,2].

We performed a preliminary evaluation of the translation quality using 500 randomly chosen sentences from travel conversation corpus. A bilingual evaluator is asked to classify the translation results into three categories: Good, Understandable and Bad. A Good sentence should have no syntactic errors and its meaning has to be correctly understood. Understandable sentences may have some errors, but the central meaning of the original sentence must be conveyed without misunderstanding. If the evaluator cannot understand the meaning of the translated result, or if it causes a misunderstanding, it is classified as Bad. With this subjective measure, the ratio of the sentences with which the meaning of the original sentence was properly conveyed, i.e. the ratio of Good or Understandable sentences, was 88% for the Japanese-to-English translation, and 90% for the English-to-Japanese translation. The ratio of Good sentences was 66% for the J-E translation and 74% for the E-J translation. This quality is equal to or better than our former system.

The total memory requirement of the whole system at the startup time is about 27MB. Requirement for the work memory is around 1 to 4MB, depending on the sentence length. For longer sentences, the translation module tends to use a larger amount of work memory.

As for the response time, a user has to wait some seconds before obtaining the recognized result when the software runs on PDAs with StrongARM 206MHz CPU. By running the software on a recent power-efficient processor such as Vr5500 (400MHz), the result is obtained almost immediately after speech input, thus realizing real-time translation. The PDAs equipped with these high-performance processors are expected to appear on the market in the near future. Other remaining issues include the microphone, robustness against background noise, and user-friendly interface.

7. CONCLUSION

We have developed an automatic speech translation system for PDAs that helps oral communication between Japanese and English speakers in various situations while traveling. Our original compact large-vocabulary continuous speech recognition engine and compact translation engine based on the lexicalized grammar led to the development of the Japanese/English bi-directional speech translation system that works with limited computational resources. Several new techniques were introduced to reduce computational cost and memory usage. The system runs on PDAs with StrongARM 206MHz CPU, 64MB RAM and a memory card. We are planning further evaluation and improvements of the system aiming at a helpful tool for cross-lingual
communication.

REFERENCES


Received January 21, 2003

Ryosuke ISOTANI received the B.E. and M.E. degrees in Mathematical Engineering and Information Physics from the University of Tokyo in 1985 and 1987, respectively. He joined NEC Corporation in 1987, and is now Principal Researcher of the Multimedia Research Laboratories. He is engaged in research on speech recognition.

Mr. Isotani is a member of the Acoustical Society of Japan, and the Institute of Electronics, Information and Communication Engineers.

Kiyoshi YAMABANA received the M.S. degree in Mathematical Science from Kyoto University in 1988. He joined NEC Corporation in 1989, and has been engaged in research of natural language processing. He is now Principal Researcher of the Multimedia Research Laboratories.

Mr. Yamabana is a member of the Information Processing Society of Japan, the Association for Natural Language Processing, the Japanese Society for Artificial Intelligence and the Association for Computing Machinery.

Shinichi ANDO received the M.E. degree in physical engineering from Osaka University in 1992. He joined NEC Corporation in 1992, and is now Assistant Manager of the Internet Solution Platform Development Division. He is engaged in research and development in the field of natural language processing.

Mr. Ando is a member of the Information Processing Society of Japan, and the Japanese Society for Artificial Intelligence.

Ken HANAZAWA received the B.E. and M.E. degrees in computer science from the Tokyo Institute of Technology in 1995 and 1997, respectively. He joined NEC Corporation in 1997, and is now working with Multimedia Research Laboratories. He is engaged in research on speech recognition.

Mr. Hanazawa is a member of the Acoustical Society of Japan, and the Institute of Electronics, Information and Communication Engineers.

Shin-ya ISHIKAWA received the M.E. degree in applied mathematics and physics from Kyoto University in 1998. He joined NEC Corporation in 1998, and is now a researcher of Multimedia Research Laboratories in Japan. He has been engaged in research and development on speech recognition, especially on the decoder for large-vocabulary continuous speech recognition.

Mr. Ishikawa is a member of the Acoustical Society of Japan.

Ken-ichi ISO received the B.S. and M.S. degrees in Physics from the University of Tokyo in 1983 and 1985. In 1986, he joined NEC Corporation, and was involved in research on automatic speech recognition. From 1991 to 1992, he was a Visiting Scientist at Carnegie Mellon University. He is now Principal Researcher of the Multimedia Research Laboratories.

Mr. Iso is a member of the Acoustical Society of Japan and the Institute of Electronics, Information and Communication Engineers.