A Speech Translation System with Mobile Wireless Clients

Kiyoshi Yamabana **Ken Hanazawa** **Ryosuke Isotani**
Seiya Osada **Akitoshi Okumura** **Takao Watanabe**
Multimedia Research Laboratories
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216-8555, Japan
{yamabana,hanazawa,isotani,osada,okumura,watanabe}@ccm.cl.nec.co.jp

Abstract

We developed a client-server speech translation system with mobile wireless clients. The system performs speech translation between English and Japanese of travel conversation and helps foreign language communication in an area where wireless LAN connection is available.

1 Introduction

Automatic speech translation has been one of prospective applications of the speech and language technology (Lavie et al., 1997; Sugaya et al., 1999; Watanabe et al. 2000). One common target application has been communication assistance for travelers in a foreign travel situation. In this situation, the user will want to carry the translation device and use it in an open space, rather than sit in a computer room in front of a display.

An obvious approach is to build a stand-alone speech translation device that is compact enough to carry around (Watanabe et al., 2000; Isotani et al., 2002). This approach is getting realistic with recent progress of hardware devices such as fast embedded CPUs and small batteries. However, because of current limitation in the amount of available computational resource, a stand-alone system has to compromise the speech translation accuracy.

In a client-server speech translation system, it is easier to provide better quality and more functionality to the users, using abundant computational resource provided by the server machines. If the situation allows use of powerful backend server machines, that is preferable to stand-alone systems.

In the Verbmobil system (Wahlster, 2000), mobile phones were used as the client device. The voice input to a mobile phone is sent to the central speech translation server as an ordinary speech data over the telephone network. A disadvantage of this approach is the narrow bandwidth of the telephone network connection. Even though an acoustic model can be trained with telephone-quality speech data, the lack of higher frequency information in the human voice affects the speech recognition accuracy. Introduction of 3G mobile phones will change this in the future, but it will take some time before they come into wider use in the world.

Recently, wireless LAN is getting popular as a means for Internet connection. Wireless LAN connection, such as Wi-Fi (802.11b), provides high-speed mobile Internet access in a limited area using standard Internet Protocol. Its speed is generally enough to transmit high-quality voice for speech recognition. It is an attractive possibility for a client-server speech translation system.

We developed a client-server speech translation system using PDAs with wireless LAN connectivity as clients. A client PDA sends a voice input to the server over a wireless LAN connection, and the server performs bidirectional speech translation between Japanese and English.

In the next section we show an overview of the system. Section 3 and section 4 describe the speech recognition module and the machine translation module, respectively. Section 5 is for discussion, and the last section concludes the paper.

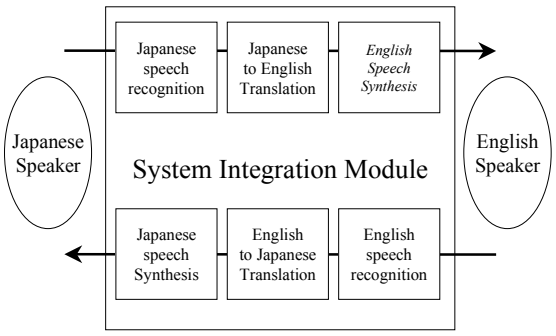

Figure 1. Overview of the Server Modules

2 System Architecture

The current system performs Japanese/English bi-directional speech translation of travel conversation. The server modules run on Windows servers, and the client modules run on Windows CE PDAs.

Figure 1 shows an overview of the server modules. It consists of Japanese and English speech recognition modules, Japanese-English bidirectional machine translation modules, Japanese and English speech synthesis modules and a system control module. The dictionary contains about 50K Japanese words and about 20K English words.

Figure 2 shows the relation of the client and the server. The client PDA provides the user interface, and the server performs speech translation processing. The client is connected to the server by TCP/IP over wireless LAN connection (802.11b). The client PDA accepts the voice from the microphone, sends it to the server, receives the translated voice and outputs it from a speaker. Multiple clients can be connected to the server simultaneously.

Input speech is converted into digital data with 16bit monaural 22 KHz sampling. Transmission of this raw data requires a speed more than 352Kbps

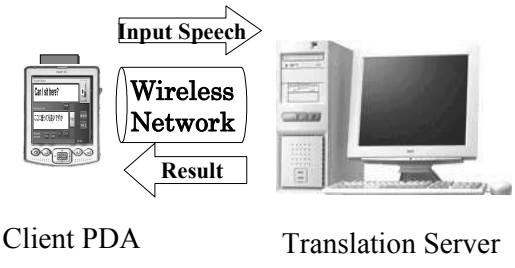

Figure 2. Relation between Client and Server

(bit per second). Although it is well below the generally available speed for wireless connection, we applied G.722 speech compression algorithm to reduce the rate to 88Kbps. The original G.722 reduces a 256Kbps stream to 64Kbps (namely 1/4), and the same core algorithm is applied to a 352Kbps stream. With a lower transmission rate, more clients can simultaneously perform speech translation task. In addition, the task becomes doable even in a less stable (slower) connection. A preliminary evaluation showed that this speech compression did not affect the speech recognition accuracy.

3 Speech Recognition Module

The speech recognition module performs large vocabulary continuous speech recognition of conversational Japanese and English based on HMMs and statistical language models. The speech recognition dictionary contains about 50K Japanese words and about 20K English words.

As shown in Figure 3, the speech recognition module consists of acoustic models, language models, word dictionaries, 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 models, the language models, and the word dictionaries are language dependent.

3.1 Acoustic Model

A speech signal is sampled at 22KHz, with MFCC analysis frame rate of 11 ms. Spectral subtraction and cepstrum mean normalization 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. 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.

3.2 Language Model

To cover the wide variety of colloquial and domain-specific expressions in travel conversation, large text corpora of travel conversation in Japanese and English has been developed and used to train the language models. The total size of the

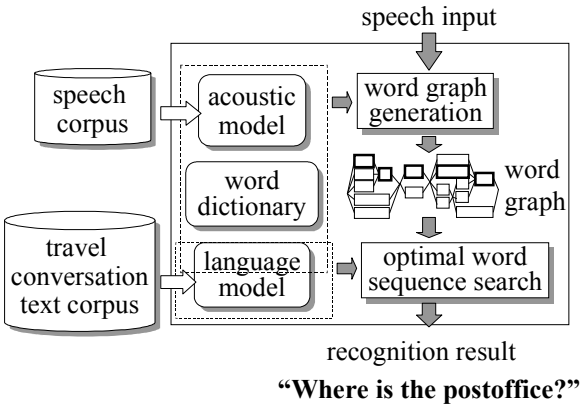

Figure 3. Speech Recognition Module

corpora is around 100K sentences for each language. The corpora contain travel conversation in various situations such as hotel, restaurant, shopping, transportation, entertainment, as well as general expressions observed in oral communication.

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

3.3 Search Engine

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

4 Machine 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.

In travel conversation, colloquial expressions are frequently used, which seldom appear in written texts such as newspapers. The 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.

With this requirement in mind, we employed the Lexicalized Tree AutoMata-based Grammar (LTAM Grammar) as the grammar formalism (Yamabana et al., 2000). This method is in line with the strong-lexicalization approach to the grammar (Schabes et al., 1988), 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 an existence of a simple bottom-up chart-parsing algorithm, which is a natural extension of the context-free grammar case.

Figure 4 is an overview of the bi-directional machine translation module. It 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. The Japanese to English translation dictionary contains about 150K words, and the English to Japanese dictionary contains about 70K words. The translation grammar has been built on generic language phenomena, and expanded for phenomena as will appear in a dialogue, such as ellipsis, idioms, fixed expressions, imperatives, requests and polite expressions.

5 Discussions

We made several experiments of the system using PC servers with Pentium 4 2GHz CPU and 512MB RAM. The whole system showed a good performance with respect to the speed and the accuracy.

We performed a preliminary evaluation of the

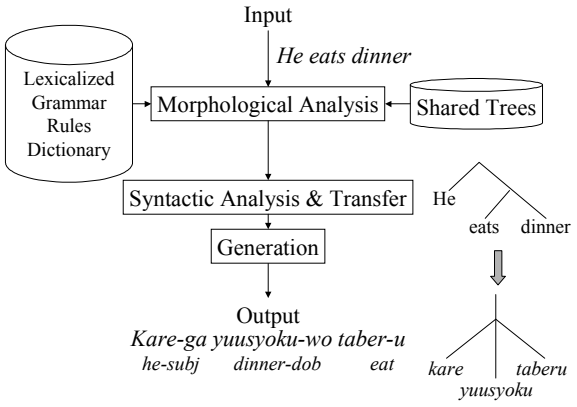

Figure 4. Machine Translation Module

speech recognition module and the machine translation module. Speech recognition accuracy was measured using 1,800 utterances by 10 Japanese male speakers and 10 English speakers, respectively. The word accuracy was 97.5% for the Japanese speakers and 92.0% for the English speakers.

A preliminary evaluation of the machine translation quality was conducted using 500 randomly-chosen sentences from travel conversation corpus. A bilingual evaluator classified 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 translation does not convey the central meaning of the original text, or if it causes a misunderstanding, it is classified as *Bad*. With this subjective measure, the ratio of *Good* or *Understandable* sentences, i.e. sentences with which the meaning of the original sentence was properly conveyed, 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 to E translation, and 74% for the E to J translation.

Compared to a stand-alone system, the developed system has better speech translation accuracy, and a quick response. In addition, future expansion to other domains and languages, for a wider use of speech translation, will be easier. An obvious disadvantage is the limitation of the area where a client can be used. However, it will not be an essential obstruction if the service area is naturally limited, for example, to the inside of a shop or a building. In this situation, the system will be able to provide a location-sensitive service, because the client location can be easily assumed with the Wireless LAN connectivity.

One problem with using PDAs as clients is an inferior quality of the built-in microphone. In the developed system, we modified the PDA hardware so that an external compact microphone is usable. This problem may disappear in the near future, since some PDAs are equipped with external microphone jacks recently. Another remaining issue is an inferior reliability of the wireless connection, which should be solved in the near future.

An ideal client will have both functions as a stand-alone speech translation device and a client device for a client-server speech translation system. The stand-alone function will be used in a usual travel situation. In a hot spot where wireless connection is available, it will behave as a client to provide better-quality location-sensitive translation. A mobile phone-based speech translation will be used to talk to a party in a distant place. Thus these functions will have respective roles depending on the situation speech translation is used.

6 Conclusion

We developed a client-server speech translation system where mobile clients are connected with the server via Wireless LAN connection. Further work includes a system design where a client operation and a stand-alone operation are compatible so that the user can use appropriate function depending on the situation.

References

R. Isotani, K. Yamabana, S. Ando, K. Hanazawa, S. Ishikawa, T. Emori, H. Hattori, A. Okumura and T. Watanabe,. 2002. “An automatic speech translation system on PDAs for travel conversation”, *Proc. ICMI-02*, pp. 211-216.

A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zahn. 1997. “JANUS III: Speech-to-speech translation in multiple languages”, *Proc. ICASSP-97*, pp. 99-102.

Y. Schabes, A. Abeillé and A. K. Joshi. 1988. “Parsing Strategies with `Lexicalized` Grammars”, *Proc. COLING’88*, pp.578-583.

F. Sugaya, T. Takezawa, A. Yokoo, and S. Yamamoto. 1999. “End-to-end evaluation in ATR-MATRIX: Speech translation system between English and Japanese”, *Proc. Eurospeech-99*, pp. 2431-2434.

W. Wahlster. 2000. “Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System”, In “Verbmobil: Foundations of Speech-to-Speech Translation”, ed. W. Wahlster, pp. 3-21, Springer.

T. Watanabe, A. Okumura, S. Sakai, K. Yamabana, S. Doi, and K. Hanazawa. 2000. “An automatic interpretation system for travel conversation”, *Proc. ICSLP-2000*, pp. IV 444-447.

K. Yamabana, S. Ando and K. Mimura. 2000. “Lexicalized Tree Automata-based Grammars for Translating Conversational Texts”, *Proc. COLING 2000*, pp 926-932.