TOWARD TRANSLATING KOREAN SPEECH INTO OTHER LANGUAGES

Jae-Woo Yang and Youngjik Lee

Electronics and Telecommunications Research Institute
Yusong P. O. Box 106, Taejon, 305-600, KOREA
E-mail: jwyang@media.etri.re.kr

ABSTRACT

This paper describes research activities of ETRI in multi-lingual spontaneous speech translation. We have developed Korean-to-English, Korean-to-Japanese speech translation system prototype that includes 5,000 word spontaneous Korean speech recognizer, Korean-English and Korean-Japanese translators, and Korean speech synthesizer with spontaneous prosody in the travel planning task. We utilize multimedia communication to increase the performance of the spoken language translation system.

1. INTRODUCTION

Spoken language translation has been a dream of speech engineers for more than three decades. The technology consists of three components, that is, speech recognition, machine translation, and speech synthesis. To accomplish a spoken language translation from one to other languages, it is essential to have all three technologies in both languages [1]. Thus, many research groups utilize international collaborations to accomplish this goal, among which the Consortium for Speech Translation Advanced Research (C-STAR) is noteworthy. Within C-STAR, each country develops its language module, and integrate the results of two countries into a bilingual translation.

ETRI joined C-STAR in 1995, and pursues various research activities in spoken language translation. This paper describes some research results and efforts of ETRI toward translating Korean speech into other languages such as English and Japanese. In section 2, we briefly introduce the research activities on system integration, and address issues on human factor, multimedia, and performance. In section 3, we describe the characteristics of Korean language, and report the status of Korean spontaneous speech database. We summarize research activities on speech recognition, machine translation, and speech synthesis in section 4, 5, and 6, respectively, and section 7 describes future works.

2. SPOKEN LANGUAGE TRANSLATION SYSTEM

To make a spoken language translation system that helps people to understand each other, the goal of spoken language translation system should not be "how to translate spoken language well" but "how to help people to communicate." This point is very critical in system design, since recognition technologies still suffer in their performance. We focus to design the system in the user's point of view, and utilize the user's intelligence as much as possible for user-to-user communication. The system should not restrict users in their position. Thus, we use microphone array for the input device, so that the user can use the system in their most comfortable position. We employ speech detection algorithm so that users just talk to the system without any control behavior. We also use multimedia user interface to increase the communication bandwidth.

Fig. 1 shows the user interface of the system. The four microphones are located in the four corners of the screen. When the user speaks, the recognized text is displayed in the dialog history window. The opponents answer is displayed in the caption window as well as the history window. The video camera located in the upper part of the screen captures the head and shoulder image of the user, and send it to the upper left window of the opponent's screen. The user and the opponent share a white board.

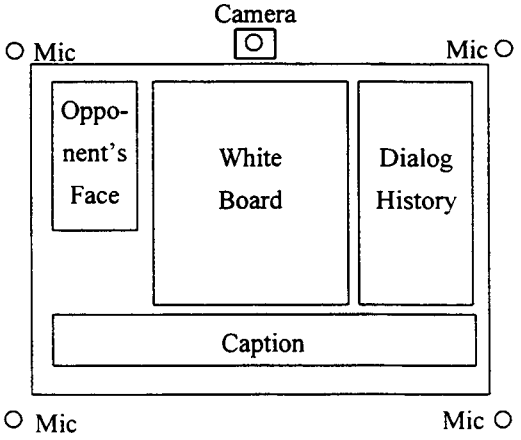

Fig. 1. Multimedia user interface of the spoken language translation system prototype.

Fig. 2 (in the last page) shows the block diagram of ETRI's spoken language translation system prototype. We use microphone array for remote speech input. To eliminate button control, we always process signal from microphone array, and

detect speech period. The speech recognizer processes detected speech signal into Korean text. To translate this text into English, we use interlingua approach. To translate the Korean text into Japanese, we use token-based transfer-driven machine translation (TB-TDMT) which will be explained in section 5. Then the English and Japanese speech synthesizer produce English or Japanese speech. We use DecTalk for English synthesizer, and KDD's synthesizer for Japanese speech synthesizer. The Korean speech synthesizer, which is not shown in Fig. 2, has spontaneous prosody.

We have proposed an end-to-end performance measure of speech translation systems [2]. Current word recognition rate and translation rate cannot measure full system performance in the user point of view. In contrast, sentence repetition rate or average time required for accomplishing a standard task can be an example of the end-to-end performance.

3. KOREAN SPOKEN LANGUAGE AND DATABASE

Korean writing system was invented in the 13th century by a group of researchers to represent Korean spoken language effectively. Each Korean alphabet represents a specific phoneme. The number of Korean alphabets is 24, and the number of phonemes is 40. A Korean syllable is a two-dimensional geometric combination of two to six Korean alphabets as in Fig. 3. There are 3,000 possible Korean syllables. A word consists of one to several syllables. Each noun usually accompanies a particle that has the same function as preposition. Each verb has variation on its tail that specifies the tense, etc. Thus, Korean is agglutinative language.

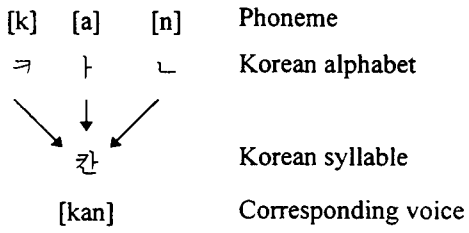

Fig. 3. The construction of a Korean syllable from Korean alphabets.

We have collected 350 Korean spontaneous dialogues in the scheduling (meeting arrangement) domain. The database has 56,149 words in 12,425 sentences. It consists of 11,083 unique words, and the bigram perplexity is 300. The Korean dialogue has much more vocabulary than English or German dialogue. There are two reasons in this phenomenon. First, Korean is agglutinative language. Thus, each noun or verb accompanies various clitics to form a word. As a result, the number of unique words becomes large. We use simple morpheme analyzer to find the stem of the words. Then the number of unique words reduces to 6,000. Second, Koreans usually spend some time to initiate a dialogue for asking recent events, chatting about the weather, etc. Such habit enlarges the number of unique words, since the topics of such chatting are out of the scheduling domain. Based on this observation, we decide to collect bilingual dialogues to prevent out-of-domain utterances.

We have collected 300 Korean-English and 300 Korean-Japanese bilingual dialogues in the travel planning task. In this case, the total number of Korean words is 28,514, and the number of unique words is 5,500. The bigram perplexity in this domain is 80.

4. KOREAN SPEECH RECOGNITION

For spontaneous speech recognition, we have developed spontaneous speech recognition platform that works in human-made noise and office environment noise based on the JANUS system. This platform works in the meeting scheduling task and the travel planning task. The prototype shows 70% word recognition rate in the travel planning bilingual dialogues whose bigram perplexity is 80.

We have developed Korean allophone clustering tree based on the acoustic-phonetic analysis of Korean phonemes. Since Korean phonemes are different from English, it is necessary to use different phoneme parameters as well as their combinations. As a result, number of Korean allophone is 3,381. We utilize this tree to train Korean spontaneous speech recognizer.

To select a best feature for recognition, we have measured the goodness of the feature vector set [3]. We define the goodness as how much the feature vectors of the same class scatter in the feature space and how much the vectors of different classes overlap. We test mel FFT, mel cepstrum, cepstrum, and wavelet feature. The mel FFT and mel cepstrum give better characteristics than the other features. We have tried to train the speech recognizer with mel-cepstrum, PLP cepstrum, and time-delayed feature. The simulation result shows that the PLP cepstrum gives the best performance. Current speech recognition system works with PLP cepstrum.

We devise an algorithm to detect long pause in spontaneous speech. When there are long pause periods in the speech, it affects the recognizer severely. Thus, we detect and delete long pause from the spontaneous speech. This increases the performance of the speech recognizer.

5. LANGUAGE TRANSLATION

The Korean and Japanese language have similar structure. Thus, proper choice of matching patterns can lead to good language translation. Such approach can operate in real time. On the other hand, the English has different language structure. Thus, it is

better to analyze the sentence into concepts, and then construct the translated sentence. With this approach, we can translate a language to another language with different structure.

We currently focus on Korean-English and Korean-Japanese speech translation. We use interlingua approach as well as token-based transfer driven machine translation (TB-TDMT) approach. We define the token as a smallest chunk of information in spontaneous speech such as a phrase. The interlingua approach is suitable for multi-lingual translation, while the TB-TDMT approach can deal with idiosynchrocy and operate in real time. In ETRI, we use interlingua approach to Korean-English translation, and transfer driven machine translation to Korean-Japanese translation.

6. SPEECH SYNTHESIS

We have developed Korean text-to-speech (TTS) conversion system using TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) synthesizer [4]. We have implemented natural prosody of read speech by adjusting pitch contour and phoneme duration [5]. Since we are implementing spontaneous speech translation system prototype, we add spontaneous style prosody by controlling pitch contour and duration of words. For this purpose, we extract pitch and phoneme duration from the spontaneous dialog. We train the prosody model using this information.

The result is very spontaneous. The spontaneity comes from the pitch contour of each words. It is interesting that the spontaneity also comes from proper pause (silence) duration and locations.

7. FUTURE WORKS

This paper describes the research status and direction of ETRI toward translating Korean speech into other languages. We have developed many new techniques in spoken language translation. Since the international collaborations in this area are very active, the system will be enhanced and demonstrated with other party. Toward this purpose, we try to integrate speech recognition and language understanding at the semantic level and improve human factors.

8. REFERENCES

1. Y. Lee, *et al.*, "Korean-Japanese speech translation system for hotel reservation - Korean front desk side," *Proc. Eurospeech'95 Madrid*, vol. 2, pp. 1197-1200, Sept. 1995.
2. J.-W. Yang, *et al.*, "Multimedia spoken language translation," *IEICE Transactions on Informatics & Systems*, June 1996
3. Y. Lee and K.-W. Hwang, "Selecting a good speech feature for recognition," *ETRI Journal*, vol. 18, pp. 29-40, Apr. 1996.
4. S.H. Kim and J.C. Lee, "Korean text-to-speech system using time-domain pitch-synchronous overlap and add method," *Proc. Fifth Aust. Int. Conf. on Speech Science Tech.*, vol. 2, pp. 587-592, Dec. 1994.
5. J.C. Lee, *et al.*, "Intonation processing for Korean TTS conversion using stylization method," *Proc. ICSPAT'95 Boston*, vol. 2, pp. 1943-1946, Oct. 1995.

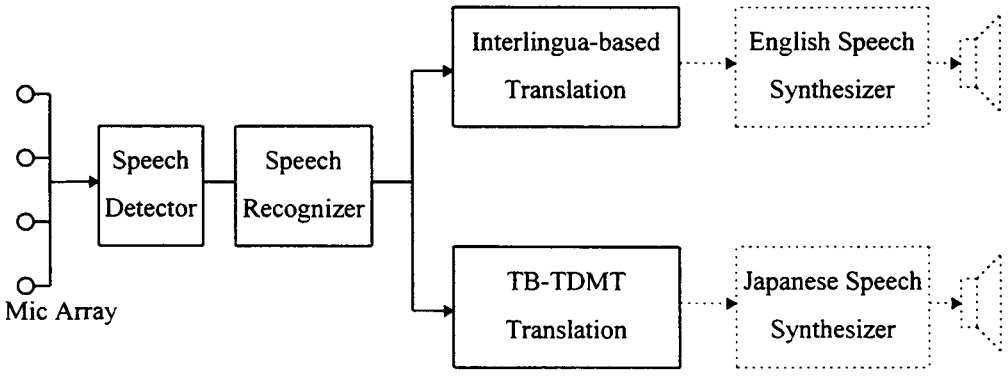

Fig. 2. System configuration of the spoken language translation system.