Abstract

This paper presents the recent advances in the Asian spoken language translation system developed by the National Institute of Information and Communications Technology/Advanced Telecommunications Research Institute International (NICT/ATR). The system was designed to translate the common spoken utterances of travel conversation from a certain source language into multi-target languages in order to facilitate a multi-party travel conversation between people speaking different Asian languages. All the system engines, including those of speech recognition, machine translation, and speech synthesis, were developed using the corpus-based approach, which statistically studied from the collection of speech and language data. Currently, all the speech-to-speech translation engines have been successfully implemented into web-servers that can be accessed by client applications worldwide. Thus, the system realizes the objective of real-time, location-free speech-to-speech translation.

1 Introduction

With the increase in globalization, international tourism, and international business, the issue of communication between people not sharing a common language has gained importance. Humankind has long dreamt of realizing a speech translation technology that will be able to break down the language barrier and thus enable instant cross-lingual communication. Many researchers have been working in the area of speech recognition, machine translation, and speech synthesis for almost five decades. With the ongoing advances in spoken language processing technology, the dream of realizing a multilingual speech translation system has become more feasible.
damental technologies are expected to be applicable to the human-machine interfaces of various telecommunications devices and services connecting Asian countries through a network using standardized communication protocols, as outlined in Fig. 1.

In this paper, we outline our contributions toward the development of a spoken language translation system catering to Asian languages within the A-STAR project. The system is designed to translate the common spoken utterances of tourists from a certain source language into multi-target languages, so that a multi-party travel conversation between tourists speaking different Asian languages can take place effectively. Currently, the system has successfully covered four Asian languages, including Japanese (Ja), Chinese (Zh), Indonesian (Id), and Vietnamese (Vi). Additionally, we also include English (En) language into the system, since it is the most widely spoken second language in the Asia region. All the system engines, including the speech recognition, machine translation, and speech synthesis system engines, were developed using the corpus-based approach, which statistically studied from the collection of speech and language data.

We first briefly provide an overview of the NICT/ATR multilingual speech-to-speech translation system architecture in Section 2. We then describe the work and experiments related to the speech recognition engine (Section 3), machine translation engine (Section 4), and speech synthesis engine (Section 5). The integration of these component technologies into a web-based speech translation system is described in Section 6. Finally, we present our summary in Section 7.

2 Overall System Architecture

Figure 2 provides an overview of the NICT/ATR multilingual speech-to-speech translation system architecture. The system has been designed for the purpose of translating spoken utterances of a certain source language into multi-target languages.

The NICT/ATR multilingual speech-to-speech translation system basically consists of three parts, namely an automatic speech recognition (ASR) engine, a machine translation (MT) engine, and a speech synthesis / text-to-speech (TTS) engine. When a person utters a Japanese sentence like “こんにちは,” the system attempts to recognize the input speech utterance through a Japanese speech recognizer. Following this, the resulting Japanese text sentence is translated into multi-target language sentences—like Chinese, Indonesian, Vietnamese, and English—by a Japanese-to-* machine translator. In this particular case, the machine gives a Chinese sentence output “Ni Hao,” an Indonesian sentence output “Selamat siang,” a Vietnamese sentence output “Xin cha’o,” and an English sentence output “Good afternoon.” Finally, all the synthesizers of these languages produce the spoken output of the resulting sentence. This translation mechanism can be used for translating any multi-party conversation comprising any or all of the Japanese, Chinese, Indonesian, Vietnamese, or English languages.

The details of the development of each component of the speech recognition, machine translation, and speech synthesis systems are described in the following section.

3 Speech Recognition Engine

3.1 Data Corpora

The training data of the Japanese, Chinese, Indonesian, Vietnamese, and English speech recognition engines are summarized in Table 1. The Japanese, English, and Chinese corpora were part of the ATR basic travel expression corpus (BTEC), while the Indonesian and Vietnamese corpora were developed by Indonesian and Vietnamese re-
Table 1: Training data of Japanese (Ja), Chinese (Ch), Indonesian (Id), Vietnamese (Vi), and English (En) speech recognition engines.

<table>
<thead>
<tr>
<th>Lang</th>
<th># Phn</th>
<th># Acc</th>
<th># Spkrs (M,F)</th>
<th># Utts</th>
<th># Hrs</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja</td>
<td>26</td>
<td>No accent</td>
<td>4200 (1600, 2600)</td>
<td>172,674</td>
<td>270.9</td>
<td>Travel expression</td>
</tr>
<tr>
<td>Zh</td>
<td>85</td>
<td>4 (BJ, SH, CT, TW)</td>
<td>536 (268, 268)</td>
<td>207,257</td>
<td>249.2</td>
<td>Travel expression</td>
</tr>
<tr>
<td>Id</td>
<td>33</td>
<td>4 (ST, JV, SN, BT)</td>
<td>400 (200, 200)</td>
<td>84,000</td>
<td>79.5</td>
<td>Daily news</td>
</tr>
<tr>
<td>Vi</td>
<td>45</td>
<td>4 No accent</td>
<td>30 (15, 15)</td>
<td>23,424</td>
<td>40.5</td>
<td>Radio broadcast</td>
</tr>
<tr>
<td>En</td>
<td>44</td>
<td>3 (US, BRT, AUS)</td>
<td>532 (266, 266)</td>
<td>207,724</td>
<td>202.0</td>
<td>Travel expression</td>
</tr>
</tbody>
</table>

search institutions, respectively.

The Indonesian corpus was developed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with ATR as a continuation of the APT (Asia Pacific Telecommunity) project (Sakti et al., 2004). This corpus contains phonetically-balanced sentences culled from the leading newspaper and magazine in Indonesia, “KOMPAS” and “TEMPO,” respectively.

The Vietnamese corpus was developed by the Institute of Information Technology (IOIT), Vietnam, under the A-STAR project. The speech was recorded from “The Voice of Vietnam” (VOV) radio programs.

In total, the data corpora consist of 4200 Japanese speakers, 536 Chinese speakers, 400 Indonesian speakers, 30 Vietnamese speakers, and 532 English speakers. Some data also covers several accents; for example, the Chinese corpus includes the following accents: Beijing (BJ), Shanghai (SH), Cantonese (CT), and Taiwanese (TW). The Indonesian corpus covers these accents: Java (JV), Sundanese (SN), and Batak (BT), and additional Standard Indonesian (ST), while the English corpus includes the following: United States (US), British (BRT), and Australian (AUS) accents.

3.2 Model Training and Evaluations

Each speech recognition engine—Japanese, Chinese, Indonesian, Vietnamese, and English—was trained separately. The experiments pertaining to all the above mentioned languages were conducted using the following feature extraction parameters: sampling frequency of 16 kHz, frame length of a 20-ms Hamming window, frame shift of 10 ms, and 25 dimensional MFCC features (12-order MFCC, ∆ MFCC, and ∆ log power).

Three states were used as the initial hidden Markov model (HMM) for each phoneme. A shared state HNet topology was then obtained using a successive state splitting (SSS) training algorithm based on the minimum description length (MDL) optimization criterion (Jitsuhiro et al., 2004). The composite multi-class bigram and trigram language models were trained using the 20K of the ATR-BTEC text sentences.

The system was tested on the ATR-BTEC test set. On an average, each language comprised about 40 speakers (20 males, 22 females), where each speaker uttered the same 510 BTEC sentences. The optimum performances of the speech recognition engines of all the languages are shown in Figure 3.

4 Machine Translation Engine

4.1 Data Corpora

The ATR-BTEC text data has functioned as the primary source for developing broad-coverage speech translation systems. The sentences forming this text data were collected by bilingual travel experts from among Japanese/English sentence pairs in travel domain “phrasebooks.” The ATR-BTEC text data has also been translated into 18 different languages that include French, German, Italian, Chinese, Korean, Indonesian, and Vietnamese. The training set of each language cor-
pus consists of 20,000 sentences while the test set comprises 510 sentences, with 16 references per sentence.

4.2 Model Training and Evaluations
The phrase-based statistical machine translation systems of the Indonesian-Japanese and Indonesian-English corpora were trained using the 20K sentence pairs of ATR-BTEC text data described above. Monolingual features—like the language model probability—were trained on the 20K monolingual text corpus of the target language.

For decoding, a multi-stack phrase-based SMT decoder called CleopATRa (Finch et al., 2007) was used. The translation quality was evaluated using the standard automatic evaluation metrics described in Table 2. The average of the respective BLEU and METEOR scores obtained for all the MT engines are listed in Tables 3 and 4, respectively.

More details about the NICT/ATR speech translation system can be found in these papers (Paul et al., 2008; Shimizu et al., 2008; Sumita et al., 2007).

Table 2: Automatic Evaluation Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>The geometric mean of the n-gram precision of the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) (Papineni et al., 2002)</td>
</tr>
<tr>
<td>METEOR</td>
<td>A metric that calculates unigram overlaps between translation and reference texts while taking into account various levels of matches (exact, stem, synonym). Scores range between 0 (worst) and 1 (best) (Banerjee and Lavie, 2005)</td>
</tr>
</tbody>
</table>

Table 3: Translation Quality of Direct Translation Approaches (BLEU)

<table>
<thead>
<tr>
<th>SRC</th>
<th>TRG</th>
<th>ja</th>
<th>zh</th>
<th>id</th>
<th>vi</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>ja</td>
<td>–</td>
<td>38.72</td>
<td>25.18</td>
<td>25.59</td>
<td>29.83</td>
<td></td>
</tr>
<tr>
<td>zh</td>
<td>43.87</td>
<td>–</td>
<td>24.85</td>
<td>24.63</td>
<td>28.12</td>
<td></td>
</tr>
<tr>
<td>id</td>
<td>32.06</td>
<td>26.42</td>
<td>–</td>
<td>41.04</td>
<td>47.01</td>
<td></td>
</tr>
<tr>
<td>vi</td>
<td>28.92</td>
<td>23.56</td>
<td>39.10</td>
<td>–</td>
<td>48.87</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>34.29</td>
<td>26.61</td>
<td>46.23</td>
<td>46.23</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Translation Quality of Direct Translation Approaches (METEOR)

<table>
<thead>
<tr>
<th>SRC</th>
<th>TRG</th>
<th>ja</th>
<th>zh</th>
<th>id</th>
<th>vi</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>ja</td>
<td>–</td>
<td>75.11</td>
<td>64.80</td>
<td>64.03</td>
<td>55.15</td>
<td></td>
</tr>
<tr>
<td>zh</td>
<td>74.06</td>
<td>–</td>
<td>65.20</td>
<td>61.74</td>
<td>54.24</td>
<td></td>
</tr>
<tr>
<td>id</td>
<td>64.92</td>
<td>65.11</td>
<td>–</td>
<td>76.07</td>
<td>68.79</td>
<td></td>
</tr>
<tr>
<td>vi</td>
<td>59.92</td>
<td>63.34</td>
<td>74.62</td>
<td>–</td>
<td>69.73</td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>65.94</td>
<td>67.22</td>
<td>80.85</td>
<td>78.49</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

5 Speech Synthesis Engine

5.1 Data Corpora
There are large corpora (Japanese, Chinese, and English) and small corpora (Indonesian and Vietnamese) that are used for the purpose of training the speech synthesis engines that were developed by NICT/ATR. They are consist as follows:

- 60 hours of Japanese female voices
- 16 hours of English male voices
- 20 hours of Chinese female voices
- 2 hours of Indonesian female voices
- 3 hours of Vietnamese male voices

5.2 Model Training and Evaluations
The development of the Japanese, Chinese, and English speech synthesis engines was based on a waveform concatenation algorithm in which appropriate subword units were selected from speech databases (Campbell and Black, 1996). Through this technique, the system could synthesize a high quality of speech.

For the Indonesian and Vietnamese languages, the training data was not large enough to build unit-concatenation speech synthesis engines. In this case, we developed statistical parametric speech synthesis systems based on the hidden Markov models (HMMs) in which speech waveforms are generated through parameters directly obtained from the HMMs (Tokuda et al., 2000). This system offers the ability to model different speech styles without the need for recording very large databases.

Here, both the Indonesian and Vietnamese statistical parametric speech synthesis systems used five state left-to-right HMMs for each phoneme-sized speech unit. These context-dependent HMMs were trained using the full contextual labels and concatenated feature vectors of extracted \( F_0 \) and mel-cepstrum parameters. The distributions for the excitation (pitch) parameter, spectral parameter, and the state duration were clustered independently using a decision-tree based context clustering technique.

The Speech waveform was synthesized using only simple excitation and the MLSA (mel-log spectrum approximation) filter (Tokuda et al., 1995). It has been observed that the Indonesian and Vietnamese speech synthesis engines are
also able to synthesize speech that resembles the speaker’s speech in the database. The speaking rate of the synthesized version is also similar to that of the natural speech case. Through informal listening tests, we have found that the synthesized speech still presents the buzziness that is characteristic of the simple excitation model. However, by and large, the prosody is good and the speech sounds smooth and stable.

6 Web-based Asian Speech Translation System

All the above spoken language translation technologies—speech recognition, machine translation, and speech synthesis engines catering to the Japanese, Chinese, Indonesian, Vietnamese, and English languages—have been successfully applied into web-servers that can be accessed by client applications. The client applications are implemented on a handheld mobile terminal device, thus realizing the objective of real-time, location-free speech-to-speech translation.

Currently, this NICT/ATR speech translation system has been connected to speech translation servers provided by other A-STAR consortium members, including ETRI (Korea), NECTEC (Thailand), BPPT (Indonesia), CDAC (India), IOIT (Vietnam), and I2R (Singapore). Therefore, all terminal devices can be connected, not only to any internal speech translation server, but also to any external speech translation server in the Asian region provided by others A-STAR members. Thus, the final speech translation system can translate not just between the Japanese, Chinese, Indonesian, Vietnamese, and English languages, but also between other major Asian languages.

7 Summary

In this paper, we have presented the NICT/ATR research activities carried out in accordance with the A-STAR consortium and conducted toward developing a multilingual speech-to-speech translation system catering to Asian languages. This system can translate source language spoken utterances into multi-target languages, so that a multi-party travel conversation between different Asian languages can take place. Currently, all the speech-to-speech translation engines have been successfully implemented into web-servers that can be accessed by client applications worldwide. The system has also been connected to other speech recognition, machine translation, and speech synthesis servers provided by other A-STAR consortium members. Thus, it realizes the objective of real-time, location-free speech-to-speech translation not only between the Japanese, Chinese, Indonesian, Vietnamese, and English languages, but also between other major Asian languages.

References


