3

Overcoming the Language Barrier with Speech Translation Technology

SATOSHI NAKAMURA
Affiliated Fellow

1 Introduction

The global, borderless economy has made it critically important for speakers of different languages to be able to communicate. Speech translation technology – being able to speak and have one's words translated automatically into the other person's language – has long been a dream of humankind. Speech translation has been selected as one of the ten technologies that will change the world. There are especially high hopes in Japan for a speech-translation system that can automatically translate one's everyday speech; the Japanese find the acquisition of foreign languages extremely difficult due to such factors as Japan's geographical conditions and the isolation of the Japanese language. As Japan and the Japanese become increasingly international, such speech-translation technology would be a great boon to the nation.

Automatic speech translation technology consists of three separate technologies: technology to recognize speech (speech recognition); technology to translate the recognized words (language translation); and technology to synthesize speech in the other person's language (speech synthesis). Recent technological advances have made automatic translation of conversational spoken Japanese, English, and Chinese for travelers practical, and consecutive translation of short, simple conversational sentences spoken one at a time has become possible (Japanese-to-English translation is equivalent to a TOEIC score of 600 or more).

There are still many challenges to overcome, however, before this technology can become viable. Some examples are the need to support more languages, and for automatic acquisition of place names, personal names, and other proper nouns. Technology should also be established for simultaneous interpretation, where a continuous stream of speech is translated. The individual technologies used in speech translation also have a wide range of applications, including speech information retrieval, interactive navigation, dictation, and summarization and archiving, and new uses are expected to emerge as well.

This report starts by affirming the significance of speech-translation technology, and providing an overview of the state of research and development to date, and the history of automatic translation technology. It goes on to describe the architecture and current performance of speech translation systems. It also touches on worldwide trends in research and development, describes the practical application of speech translation technology, its extension to Asian languages, and describes efforts to standardize interfaces. It concludes by summarizing the challenges and future prospects for speech-translation technology, and proposes a way forward to resolve these challenges in the promotion of this technology.

2 The history of speech translation technology

2-1 The significance of speech translation research and history to date

Speech translation is a technology that translates spoken language into speech in another language. Speech-translation technology is significant because it enables speakers of different languages from around the world to communicate, erasing the language divide in global business and cross-cultural exchange. Achieving speech translation would have tremendous scientific, cultural, and economic value for humankind. The article "10 Emerging Technologies

That Will Change Your World" in the February 2004 issue of An MIT Enterprise Technology Review lists "Universal Translation" as one of these ten technologies. The article showcases a number of translation technologies, but focuses on speech-translation technology.

Speech translation first grabbed attention at the 1983 ITU Telecom World (Telecom '83), when NEC Corporation performed a demonstration of speech translation as a concept exhibit. Recognizing that many years of basic research would be required to implement speech translation, the Advanced Telecommunications Research Institute International (ATR) was subsequently founded in 1986, and began a project to research speech translation. Researchers from a wide range of research institutes both in Japan and internationally joined this project[1]. In 1993, an experiment in speech translation was conducted linking three sites around the world: the ATR, Carnegie Melon University (CMU), and Siemens. After the start of ATR's project, speech translation projects were started around the world. Germany launched the Verbmobil project; the European Union the Nespole! and TC-Star projects; and the United States launched the TransTac and GALE projects. The GALE project was started in 2006 to translate automatically Arabic and Chinese into English. The goal of this project is to automate the extraction of vital multilingual information that up until then had been performed by humans; the project architecture consists of a batch text-output system. In contrast, the objectives of the ATR and NEC are speech translation enabling face-to-face and non-face-to-face cross-language communication in real time. Online speech-to-speech translation is thus an integral component of this research, and immediacy of processing is a key factor.

Speech translation integrates three components: speech recognition, language translation, and speech synthesis. Each of these technologies presents its own difficulties. In particular, a requirement of this technology is recognizing and translating spoken language; this is much more difficult than translating text, because spoken language contains ungrammatical, colloquial expressions, and because it does not include punctuation like question marks, exclamation marks, or quotation marks. Mistakes in speech recognition also cause major translation errors. Consequently, researchers have chosen a development technique of increasing accuracy to a usable level by initially restricting the system to relatively simple conversation, rather than supporting all forms of conversation from the start. Table 1 shows the history of speech-translation technology. Research and development has gradually progressed from relatively simple to more advanced translation, progressing from scheduling meetings, to hotel reservations, to travel conversation. Moving forward, however, there is a need to further expand the supported fields to include a wide range of everyday conversation and sophisticated business conversation.

Table 1 : Trends in the Research and Development of Speech Translation

Research Phase	1980s Confirmation of Feasibility	1990s Extension of Technology	2000s Attempts at Practical Systems
Fields	Simple reservations (ATR-phase 1)	Reservations and scheduling (ATR-phase 2, Verbmobil)	*Everyday travel conversation (ATR-phase 3) *Translation of keynote speeches (TC-Star) *Conversation for military use (TranTac) *Intelligence collection (Gale)
Linguistic features	Grammatically correct expressions	Everyday expressions that may be context-dependent or ungrammatical	Expressions including a wide range of topics and proper nouns
Phonological features	Clear pronunciation	Unclear pronunciation	Audio including background noise
Translation method	Rule-based translation Translation using artificial intermediate language	Example-based translation Translation using English as intermediate language	Statistically-based translation Direct translation of multiple languages

Note: ATR-phase 1: 1986 to 1992; ATR-phase 2: 1993 to 1999; ATR-phase 3: 2000 to 2005. For other projects, refer to the text.

Prepared by the STFC

2-2 The history of automatic translation

Of the three components of speech translation, recent advances in text translation technology have made a major contribution to the realization of automatic speech translation technology. Research into text translation technology has a long history going back more than half a century.

Warren Weaver of the Rockefeller Foundation advocated research into automatic translation technology for text in 1946, shortly after the birth of the first computer. At the time, the Rockefeller Foundation had a huge influence on the United States' science and technology policy. Then in 1953, Georgetown University and IBM began joint research of automatic translation using the 701 (the first commercial computer, developed by IBM). In 1954, the world's first automatic translation system was built on this computer, demonstrating the possibility of translation from Russian to English. Consisting of a dictionary of 250 terms and 6 rules, the translation capabilities of this system were extremely limited, but the demonstration had a huge impact on society. People at the time felt that the language barrier would soon be knocked down. Subsequently, as part of its response to the shock of the Sputnik launch, a whopping $20 million were invested into research on automatic translation by US government.

In 1965, however, the Automatic Language Processing Advisory Committee (ALPAC) presented a grave report to the US National Academy of Sciences. The report stated that because automatic translation would not be practical for the foreseeable future, research efforts should instead be directed at language theory and understanding to serve as the underpinnings of this technology. In the US, budgets for automatic translation were subsequently cut, and the focus turned to basic research, with the key concepts being meaning and understanding. One famous result from this time is language understanding using world knowledge by Winograd in 1970. The base knowledge base in this kind of research, however, was insufficient, and it cannot be said to have tied directly into improved performance of automatic translation in a general or practical sense.

Three great waves of technological innovation hit Japan in the 1980s: rule-based translation, example-based translation, and statistically-based translation. In Japan, a project to translate abstracts of the science and technology literature of the Science and Technology Agency (dubbed the Mu project) was successful. As a result, research and development into rule-based automatic translation, based on dictionaries and rules (analytic grammar rules, conversion rules, and generative grammar rules), began to gain popularity. A venture named Bravis launched sales of a commercial translation program. This spurred the commercialization of automatic-translation software by such big-name IT companies as Fujitsu, Toshiba, NEC, and Oki Electric Industry. All of the commercial software packages in the world today, and nearly all of the Web-based software, have this rule-based technology as their cores. Because better and more complete specialized dictionaries were an effective way to improve translation quality, slow but steady efforts have built up to increase dictionary sizes from a few tens of thousands of entries to millions of entries.

Meanwhile, in 1981 professor Makoto Nagao of Kyoto University took a hint from the translation process carried out by humans to propose an example-based translation method using sentences similar to the input sentence and their translations (together called "example-based translations"). This example-based translation, combined with further research at Kyoto University and ATR around 1990, created a second wave that spread from Japan to the rest of the world. This method has been incorporated into some commercial rule-based systems; it is also currently being used as the core method for a Japanese-to-Chinese translation project for scientific and technical publications being led by the National Institute of Information and Communications Technology (NICT).

Then in 1988, IBM proposed a method called statistical machine translation, combining pure statistical processing that excludes grammatical and other knowledge with a bilingual corpus. This method did not get attention for some time, however, for a number of reasons: the paper was difficult to understand, computer performance was lacking, the translation corpora were too small, the method of execution was only published in patent specifications, and it was not effective for languages other than related languages like English and French. Around 2000, however, a new method called phrase-based statistical machine translation was proposed, and buoyed by more complete bilingual corpora and more powerful computers, this created the third major wave.

Today, nine out of ten research papers in the field are on statistically-based translation. It is difficult to tell at this time whether this research domain will continue to grow.

Today, the three waves above are just now overlapping. We have gradually come to learn the strengths and weaknesses of the rule-based, example-based, and statistically-based approaches to automatic translation. The current opinion is that the best performance can be achieved by fusing these three approaches in some way, rather than by using any one of them in isolation. The three methods, however, have a common problem: they all translate at the sentence level. They cannot use contextual information. In other words, they do not make use of the relationships with the surrounding text, and thus cannot ensure cohesion. Statistical machine translation in particular performs automatic translation without analyzing the meaning of the input sentence, and so sometimes generates nonsensical translations.

The method of using example-based and statistically-based methods is called "corpus-based translation,"and this paper primarily presents methods using statistical machine translation. A corpus is a database of text with supplementary linguistic information added, such as pronunciations, part-of-speech information, and dependency information. The next and subsequent chapters primarily describe corpus-based translation methods.

3 Overview of speech translation technology and performance

3-1 Multilingual speech translation processing architecture

Figure 1 shows the overall architecture of the speech-translation system. Figure 1 illustrates an example where a spoken Japanese utterance is recognized and converted into Japanese text; this is then translated into English text, which is synthesized into English speech. The multilingual speech-recognition module compares the input speech with a phonological model consisting of a large quantity of speech data from many speakers (the model consists of the individual phonemes making up the speech utterances), and then converts the input speech into a string of phonemes represented in the Japanese katakana syllabary. Next, this string of phonemes is converted into a string of words written in the Japanese writing system (mixed kana and kanji characters), so that the probability of the string of words is maximized. In this conversion, string of words appropriate as a Japanese utterance is generated based on the occurrence probability of a string of three words using an engine trained on large quantities of Japanese text. These words are then translated by a conversational-language translation module, replacing each Japanese word in the string with the appropriately

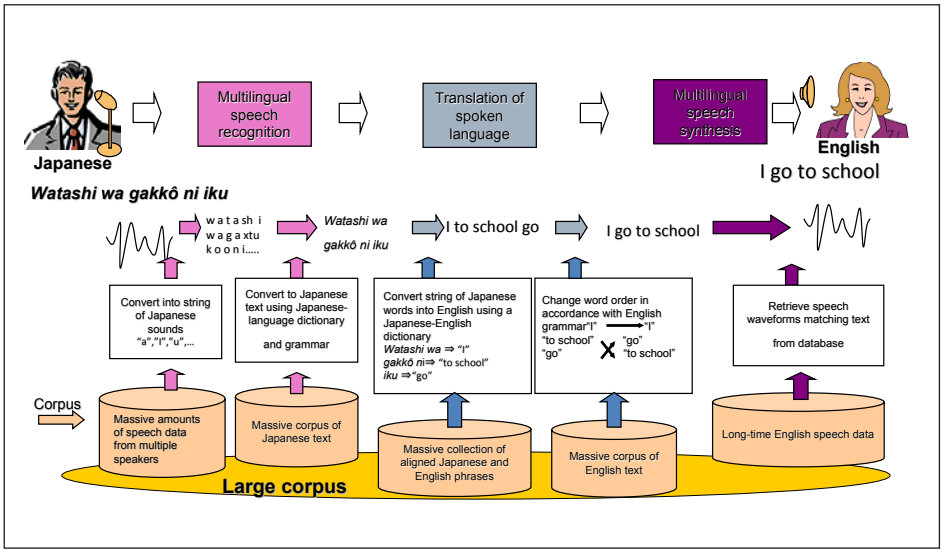

Figure 1 : The mechanism of the speech translation system

Prepared by the STFC

corresponding English word. The order of the English words is then changed. In this procedure, the Japanese words in the string are replaced by English words using a translation model trained on pairs of Japanese-English translations. In order to rearrange the words into a proper English utterance, a string of words appropriate as an English utterance is generated based on the occurrence probability of a string of three words using an engine trained on large quantities of English text. This is then sent to the speech synthesis module. The speech synthesis module estimates the pronunciation and intonation matching the string of English words, selects matching waveforms from a database of long-time speech data, connects them, and performs high-quality speech synthesis. The method of speech recognition and synthesis using statistical modeling and machine learning based on massive speech corpora is called "corpus-based speech recognition and synthesis."

ATR developed its speech-translation system[1,2] by collecting a corpus of general spoken travel conversation, in order to implement speech translation of travel conversation. To date, the project has created a Basic Travel Expression Corpus (BTEC) consisting of 1,000,000 matched pairs of Japanese and English sentences, and 500,000 each of matched Japanese-Chinese and Japanese-Korean pairs. This is the world's largest translation corpus of multilingual travel conversation. The English sentences in the corpus are an average of seven words long and cover such everyday travel-conversation topics as greetings, problems, shopping, transportation, lodging, sightseeing, dining, communication, airports, and business. Below is an example of spoken English translations of a Japanese sentence.

The Japanese is "mado o akete mo ii desu ka"
Here are the corresponding English sentences:

1. may i open the window
2. ok if i open the window
3. can i open the window
4. could we crack the window
5. is it okay if i open the window
6. would you mind if i opened the window
7. is it okay to open the window
8. do you mind if i open the window
9. would it be all right to open the window
10. i'd like to open the window

As can be seen by these examples, the utterances used in speech translation are not complete sentences – they often lack subjects, and no capitalization is used in subjects and proper nouns – and even questions do not have question marks. It is also necessary to handle extremely colloquial expressions.

In addition to BTEC, data collected from a corpus of about 10,000 utterances of dialog recorded under real-life conditions mediated by a speech translation system called Machine Aided Data (MAD) was evaluated; also evaluated was a dataset called Field Experiment Data (FED). This data was collected via a field experiment performed at Kansai International Airport, with the cooperation of the prefecture of Osaka, over a total of five days between December 2004 and January 2005. The data consists of a total of about 2,000 utterances of conversation between foreign-language speakers (39 English speakers and 36 Chinese speakers) and guides at a tourist center mediated by the speech-translation system.

3-2 Comparative study with human speech-translation capability

It is extremely difficult theoretically to evaluate the accuracy of speech translation. If the evaluation of the speech synthesis module is not included, evaluation is made by feeding a number of test sentences into the system, and evaluating the quality of the output. In this sense, the method for evaluating speech translation is essentially the same as that for evaluating automatic text translation. For speech translation, however, the utterances that are evaluated are not strings of text but speech.

Two methods are used to evaluate translation quality: one method where the translations are manually given subjective ratings on a five-point scale, and another that compares the similarity between the output of the system and previously prepared reference translations. A number of rating scales have been proposed for the latter, including BLEU, NIST, and word error rate (WER). Recently, these scales have come to be widely used[4]. Since these results are simple numerical values, it is possible to use them to compare two different systems. What these scores cannot answer, however, is how the system with the higher score will perform in the real world.

A method has been proposed to resolve this issue, by estimating system performance in human terms, estimating the system's corresponding Test of English

for International Communication (TOEIC) score. First, native speakers of Japanese with known TOEIC scores ("TOEIC takers") listen to test Japanese sentences, and are asked to translate them into spoken English. Next, the translations by the TOEIC takers are compared against the output of the speech-translation system by Japanese-English bilingual evaluators. The human win rate is then calculated as the proportion of tests sentences for which the humans' translations are better. After the human win rate has been completely calculated for all TOEIC takers, regression analysis is used to calculate the TOEIC score of the speech-translation system. Figure 2 shows system performance converted into TOEIC scores. When using relatively short utterances like those in basic travel conversation (BTEC), the speech-translation system is nearly always accurate. The performance of the speech-translation system on conversational speech (MAD and FED) is, however, equivalent to the score of 600 (TOEIC) by the Japanese speakers.

Furthermore, performance drops significantly when dealing with long, rare, or complex utterances. There is thus still room for improvement in performance.

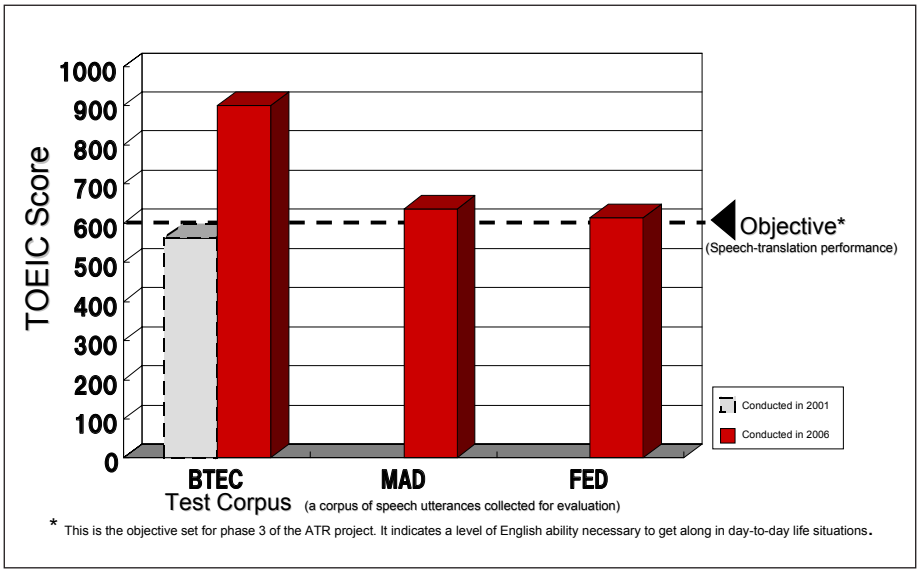

Figure 2 : Example of evaluating the accuracy of speech translation using TOEIC scores

Source : Reference [1]

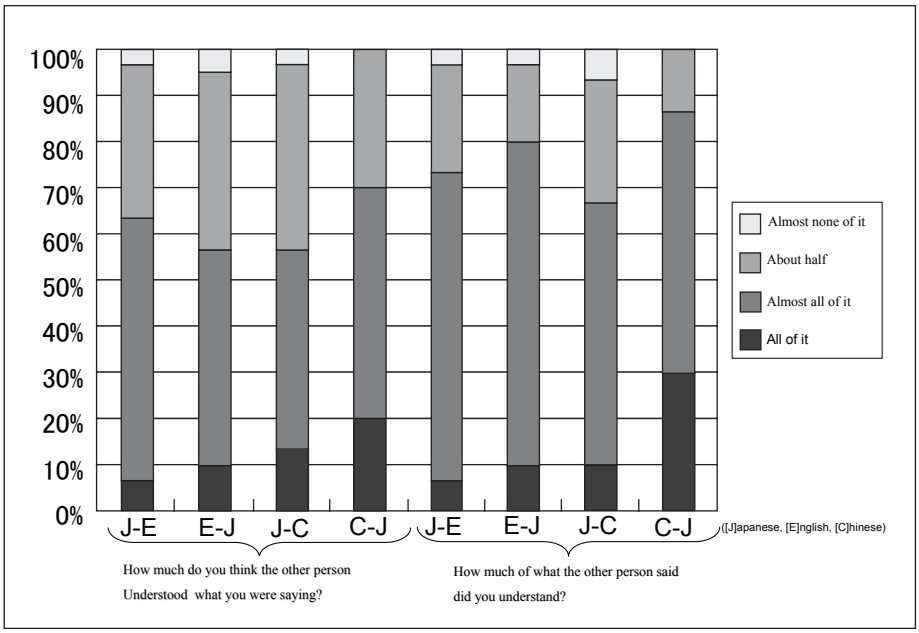

Figure 3 : Evaluation of level of comprehension based on questionnaires

Source : Reference [6]

3-3 Field experiments using speech translation device

A field experiment was conducted in Downtown Kyoto from 30 July to 24 August 2007[6], with the objective of evaluating the characteristics of communication mediated by a stand-alone speech translation device about the size of a personal organizer, as well as evaluate the usability of this device. The field experiment was set up as follows, in order to minimize the restrictions on the test subjects: (1) The people with whom the subjects conversed were not selected ahead of time, in order to collect a diverse range of expressions while using the speech-translation device in realistic travel situations, such as transportation, shopping, and dining. (2) Although the subjects were told the purpose of the dialog ahead of time, no restrictions were placed on the exact destination or proper names of items to purchase. (3) Subjects were allowed to change the topic freely depending on the flow of the conversation. (4) Subjects were allowed to move to different locations as appropriate, in accordance with the task. (5) No time limit was placed on single dialogs.

In the case of transportation, the objective was considered to have been met if the subject was able to obtain information about the destination or to actually travel there. For shopping and dining, the objective was met if the subject completed the purchase of the article or the meal and received a receipt.

In addition to quantitative evaluations of speech recognition rates, dialog response rates, and translation rates, the experiment also evaluated the level of understanding based on questionnaires. As shown in Figure 3, in the evaluation of the level of understanding of 50 native English speakers, about 80% said that the other person understood nearly everything that they said, and over 80% said they understood at least half of what the other person said. This result suggests that the performance of speech-translation devices could be sufficient for communication.

4 Worldwide trends in research and development

International evaluation workshops give a strong boost to the development of speech-translation technologies. An international evaluation workshop is a kind of contest: the organizers provide a common dataset, and the research institutes participating in the workshop compete, creating systems that are quantitatively evaluated. The strengths and weaknesses of the various proposed algorithms are rated from the results of the evaluation, and the top algorithms are then widely used in subsequent research and development. This allows research institutes to perform research both competitively and cooperatively, promoting efficient research. Some representative examples of international evaluation workshops are presented here, describing automatic evaluation technologies that support competitive research styles via evaluation workshops.

(a) The International Workshop on Spoken Language Translation (IWSLT)[7] is organized by C-STAR, an international consortium for research on speech translation including ATR in Japan, CMU in the United States, the Institute for Research in Science and Technology (IRST) in Italy, the Chinese Academy of Sciences (CAS), and the Electronics and Telecommunications Research Institute (ETRI) in Korea. The workshop has been held since 2004. Every year, the number of participating institutes increases, and it has become a core event for speech translation research. The subject of the workshop is speech translation of travel conversation from Japanese, Chinese, Spanish, Italian, and other languages into English. Two distinguishing features of the IWSLT are that it is for peaceful uses (travel conversation) and that the accuracy of the translation is fairly good, because it is a compact task.

(b) Global Autonomous Language Exploitation (GALE) [8] is a project of the US Defense Advanced Research Projects Agency (DARPA). It is closed and non-public. US $50 million are invested into the project per year. The purpose of the project is to translate Arabic and Chinese text and speech into English and extract intelligence from them. A large number of institutions are divided into three teams and compete over performance. The teams are run in units of the fiscal year in which the targets are assigned, and every year the performance is evaluated by outside institutions. In the United States, research on automatic translation is currently strongly dependent on DARPA budgets, and the inclinations of the US Department of Defense are strongly reflected.

Methods for evaluating translation quality have

become a major point of debate at these workshops. There are various perspectives on translation quality, such as fluency and adequacy, and it has been considered a highly knowledge-intensive task. A recently proposed evaluation method called BLEU is able to automatically calculate evaluation scores with a high degree of correlation to subjective evaluations by humans. This makes it possible to develop and evaluate systems repeatedly in short cycles, without costing time or money, which has made translation research and development much more efficient[4].

5 Practical applications of speech translation technology

The improved processing power and larger memories of computers and more widespread networks are beginning to make it possible to implement portable speech translation devices. Advances are being made in the development of standalone implementations in compact hardware, and distributed implementations connecting mobile phones and other devices to high-performance servers over a network.

It is not feasible to implement the standalone method on a computer that is carried around, due to such issues as size, weight, and battery lifetime. There is also expected to be demand in situations where wireless and other infrastructure is not available. In light of these issues, efforts are being directed toward the commercialization of dedicated mobile devices with built-in speech-translation functionality. In 2006, NEC developed the world's first commercial mobile device (with hardware specifications of a 400-MHz MPU and 64 MB of RAM) with onboard Japanese-to-English speech translation.

Meanwhile, in November 2007 ATR developed a speech translation system for the DoCoMo 905i series of mobile phones as a distributed implementation using mobile phones and network servers. The system, called "shabette honyaku" (see Figure 4), was released by ATR-Trek, and is the world's first speech translation service using a mobile phone. Then in May 2008, a Japanese-to-Chinese speech-translation service was begun on the DoCoMo 906i series. Figure 5 shows the architecture of the speech recognition module used in the distributed speech translation. The mobile phone (front end) performs background noise suppression, acoustic analysis, and ETSIES 202 050-compliant encoding[9], and sends only the bit-stream data to the speech recognition server. The speech recognition server (back end) then expands the received bit-stream, performs speech recognition, and calculates word reliability. One of the benefits of using this system architecture is that it is not bound by the information-processing limitations of the mobile phone, making large-scale, highly precise phonological and linguistic models to be used. Since these models are on the server and not the mobile phone, they are easy to update, making it possible to keep them up to date at all times. The system is already in wide use: as of June 2008, there have been a cumulative total of over five million accesses.

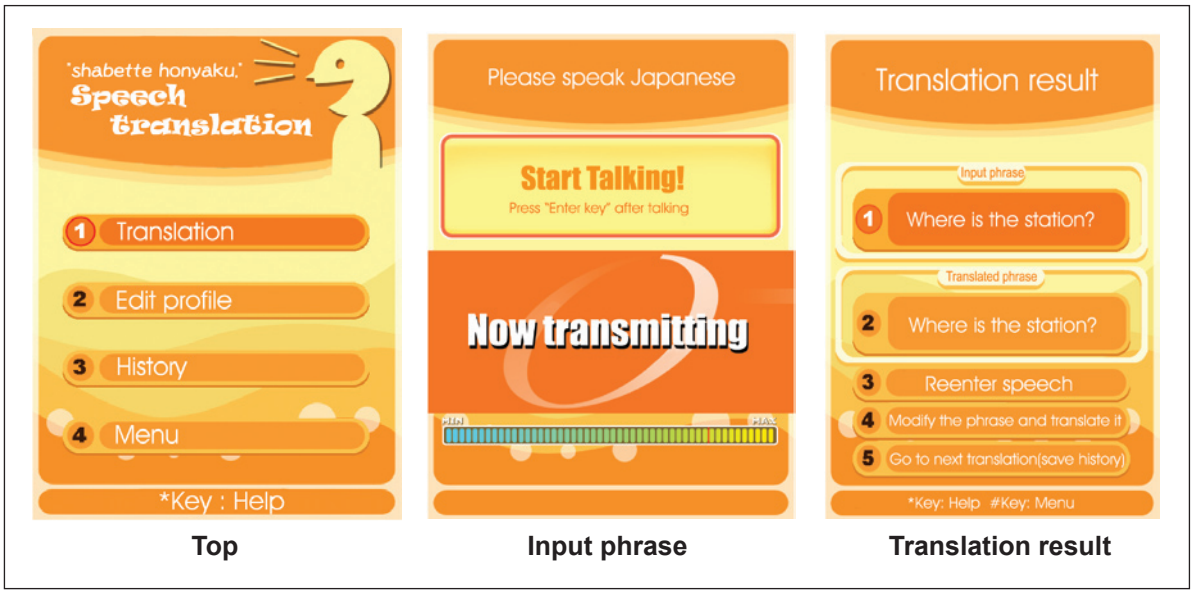

Figure 4 : The world's first speech translation service using a mobile phone

Source : ATR-Trek Co., Ltd.

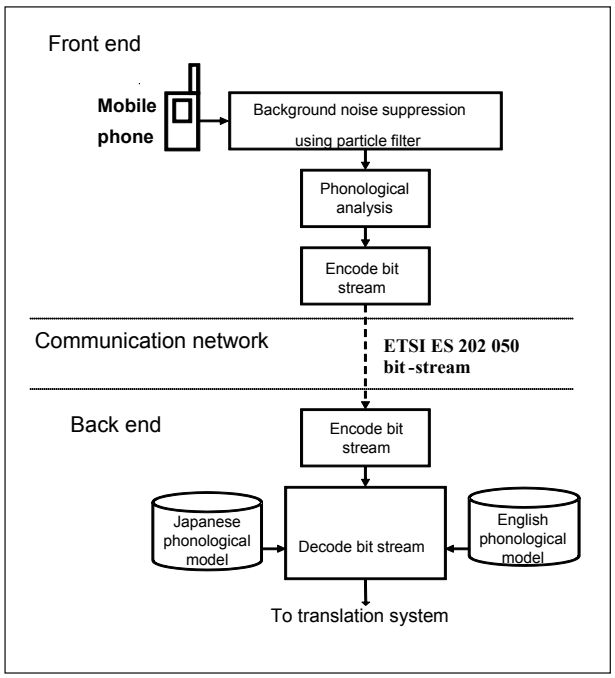

Figure 5 : Architecture of speech recognition module used in distributed speech translation

Prepared by the STFC based on the references [9,12]

6 Standardization for support of multiple languages in speech translation

As speech translation technology overcomes linguistic barriers, it would be preferable for researchers and research institutions from many different countries to research it jointly. The C-STAR international consortium for joint research of speech translation, in which ATR and CMU play a central role, has been quite active in international joint research.

Meanwhile, the foreign travel destinations of Japanese people – whether for tourism, emigration, or study abroad – are becoming more diverse, and people from a large number of countries are coming to Japan in increasing numbers for tourism, study, and employment. These and other changes are heightening the need for means of interaction with people from non-English speaking countries.

In particular, Japan is strengthening its social and economic ties in the Asian region including Russia, and enhancing mutual understanding and economic relations at the grassroots level has become a key challenge. Relations with the rest of Asia are more vital to Japan than ever before. Consequently, rather than English, Japan needs to be able to get along in the languages of its neighbors, such as Chinese, Korean, Indonesia, Thai, Vietnamese, and Russian – languages that until now have not been widely taught or spoken in this country.

Against this backdrop, A-STAR was founded as a speech translation consortium for creating the basic infrastructure for spoken language communication overcoming the language barriers in the Asia-Pacific region. Rather than the research and development of technology proper, however, the consortium's objective is to establish an international joint-research organization to design formats of bilingual corpora that are essential to advance the research and development of this technology, to design and compile

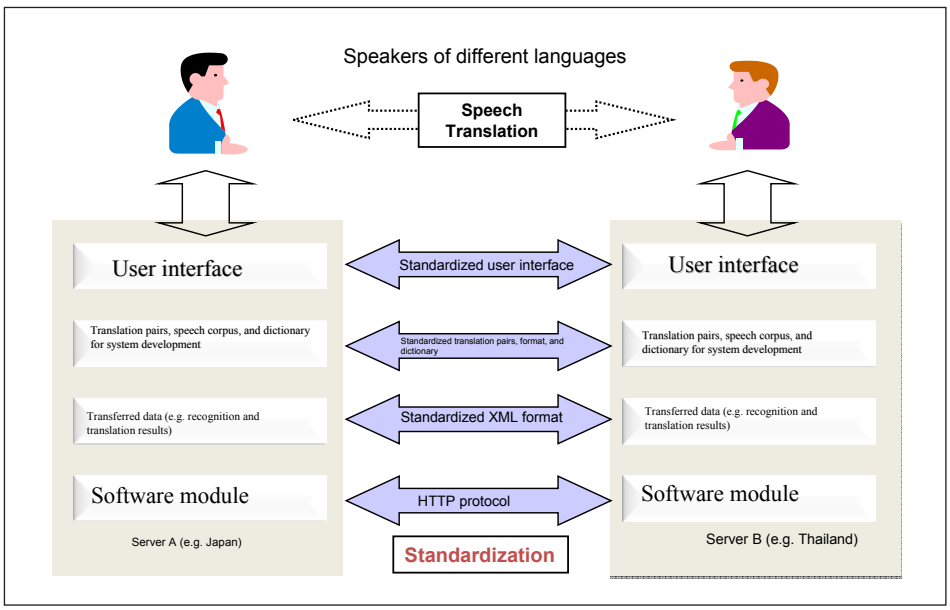

Figure 6 : Illustration of speech translation standardization

Source : Reference [12]

basic bilingual corpora between Asian languages, and to standardize interfaces and data formats to connect speech translation modules internationally, jointly with research institutions working in this field in the Asia-Pacific region. The consortium's activities are contracted as research by the Asia Science and Technology Cooperation Promotion Strategy, which is a project of the special coordination funds for promoting science and technology. This project has further been proposed and adopted as APEC TEL (Telecommunications and Information)[10] project. It is also moving to create an expert group in the APT ASTAP (Asia-Pacific Telecommunity Standardization Program) in order to create a draft of the standardized interface and data formats for connecting speech-translation modules[11]. Figure 6 illustrates the standardized connections being considered in this project. This will standardize the interfaces and data formats of the modules making up the speech translation architecture, in order to enable their connection over the Internet. It is also necessary to create common speech-recognition and translation dictionaries, and compile standardized bilingual corpora. The basic communication interface will be Web-based HTTP 1.1 communication, and a markup language called STML (speech translation markup language) is currently being developed as the data format for connecting applications[12].

7 Challenges and future prospects of speech translation technology

7-1 Challenges for the development of speech translation

As described above, speech translation is a technology enabling communication between speakers of different languages. There are still many research challenges to overcome, however: in particular, there is great speaker dependency and diversity of expression; additionally, new words and concepts are constantly being created in accordance with changes in society. Speech translation technology is currently at the level of simple utterances of about seven words in length, such as travel conversation. Consequently, there are still many unsolved challenges before speech translation will be capable of handling long, complex speech such as a newspaper or lecture. Below are listed some of the immediate technical challenges.

1) Evaluating and Improving Usability in Practical Applications

Human speakers have many inherent differences. People have many differences in speaking style, accent, and form of expression. Speech translation must aim to suppress variations in performance due to these differences, and provide the same high level of performance for all users. Additionally, acoustic noise, reverberation, and speech by other speakers have a huge impact during real world use. Measures to remedy these external factors are also extremely vital. Meanwhile, from the standpoint of usability as a communication tool, it is essential to further reduce the time from speech recognition to translation to speech synthesis. When speech translation is used, the user does not understand the translation language. For this reason, there are no techniques to check whether a translation is correct. A method must thus be provided for the user to check whether the translation is correct, by such means as translating it again back into the user's language, or back-translating it. When considering it as a tool for gathering information while traveling, it is also essential to at the same time provide a means to gather information via the Internet in multiple languages, not only by asking people.

These challenges require field testing and technology development to be performed in parallel, as well as a growth loop of data collection, improving performance, improving usability, and providing trial service.

2) Support for multiple languages

Although English is becoming the de facto worldwide lingua franca, what is needed is not a system that will translate into only English, but one that will translate directly into the 6,000 languages said to exist on our planet today. Multilingual speech translation requires a system of speech recognition, translation, and speech synthesis for each of these languages. In other words, massive speech corpora, bilingual corpora, and text corpora are required for each of these languages. The collection of speech corpora in particular is extremely expensive. This type of technology could also have great value in the sense of preserving languages in a process of decline and extinction.

3) Standardization for the connection of speech translation worldwide via the network

Module connections are also being standardized in the Asia-Pacific region. Moving forward, it will

be necessary to advance standardization for wide international connectivity, and the development of a joint-research structure.

4) Relaxing of copyright to enable example translations to be used via the web

The development of speech translation technology requires a text corpus of the source language, a text corpus of the translation language, a bilingual corpus of translations between the two languages, and speech corpora. It is extremely expensive to create and collect these corpora using conventional methods. One method that is currently gaining attention is collecting data from the Web via the Internet, which continues its explosive growth. For example, the secondary use of news and other media published in multiple languages would be an effective way to improve the performance of speech translation. As of this time, however, copyright issues have not been resolved.

5) Using the latest proper nouns based on the user's current location

There are huge numbers of proper names of people, places, and things. Incorporating all of these proper nouns into the speech-translation system at the same time would be nearly impossible, both in terms of performance and time. It would therefore be efficient to automatically acquire proper nouns corresponding to the user's location using GPS or the like, and perform speech recognition, translation, and speech synthesis tailored to that location.

7-2 Research and development roadmap

Figure 7 shows the history of speech translation to date, and indicates future directions of research and development. In 2010, an international research consortium on Asian languages plans to prototype speech translation via the Internet.

It is conjectured that the international research consortium will come out with a prototype including Western European languages and with greater standardization of interfaces by around 2015. Japan's Project to Accelerate Benefits to Society (described in the next section) plans to establish technology for networked speech translation by 2012, after various field testing. In the mid to long term, speech translation capable of continuous simultaneous interpreting of business and lectures is expected to be available by around 2015, and by 2025, multilingual simultaneous interpretation is expected to be available that is capable of contextual awareness and summarization, gradually bringing us closer to the dream of simultaneous interpretation.

8 Japanese policy regarding speech translation

8-1 The cabinet office project to accelerate benefits to society

"Creating spoken-communication technology to overcome the language barrier"

The speech translation project of the Japanese Ministry of Internal Affairs and Communications has been set as the Cabinet Office Project to Accelerate Benefits to Society. This project has been active since fiscal 2008. Amidst increasing globalization, the goal of the project is to deepen mutual understanding between nations through direct international communication at the level of individual citizens. The project aims to create an automatic speech translation system enabling Japanese people to break the language barrier and speak and interact directly with people from other countries in the Asia-Pacific region and elsewhere. Giving consideration to current user needs and the technological advances that can be expected over the next five years or so, the project is planning and advancing field testing of such applications as tourism, shopping, and international-exchange events. After the project ends, it will quickly be tied into commercial services in industry, accelerating the benefit to society derived from the results of the project.

The project is developing technology to distribute proper nouns like place names and personal names, as well as translation knowledge corresponding to a wide range of topics, over the network. It is working to establish network-based speech-translation technology combining this network with translation devices. It is also working to make innovations in speech-translation communication more visible, in order to help popularize it and facilitate the advent of practical/commercial systems. This effort to make the technology more visible is based on the awareness that the mismatch between developers and users inhibits practical/commercial applications. It thus aims to publicize progress in technology development as appropriate, in order to enable users to envision themselves using it (e.g. what kind of situations they

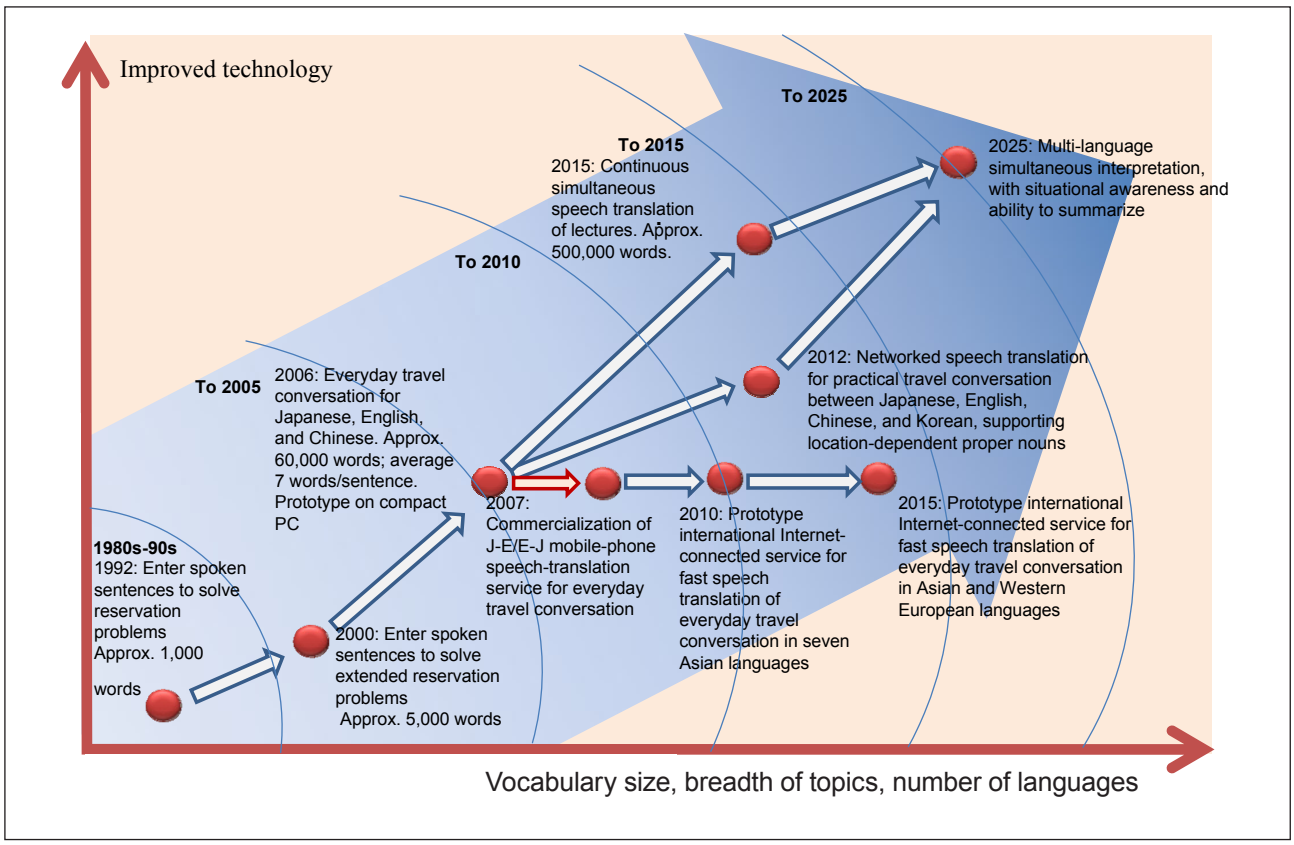

Figure 7 : Trend forecast for research and development of speech translation technology

Prepared by the STFC

could use it in). It will then build functionality and user interfaces suited to these types of use, and repeatedly field test it in concrete situations. By repeatedly validating the technology through close collaboration with both developers and users, the project will play a vital role in a grassroots international exchange by enabling ordinary travelers to enjoy overseas travel with almost no inconveniences in Japan and the English and Chinese-speaking worlds. The project is also expected to create new business and promote local industry and development.

8-2 Special ubiquitous zone

The Special Ubiquitous Zone Project was created based on the Ministry of Internal Affairs and Communications' "ICT Innovation Promotion Program" and "ICT International Competitiveness Enhancement Program." Planned to run for three years starting in fiscal 2008, the project's aim is to support the development and field testing of innovative services. The Special Ubiquitous Zone project was adopted from a proposal by a consortium of eight corporations (Intage Inc.; Toei Kyoto Studio Co., Ltd.; NICT; ATR; JTB Business World Tokyo Corp.; Willcom, Inc.; and NEC), represented by Kyoto Industrial Support Organization 21. This project is being conducted in collaboration with the prefecture of Kyoto, to develop a mobile-phone service to provide market surveys, multilingual translation, and tourist information targeting foreign visitors to Kyoto, and to field test ubiquitous multifunctional servers supporting next generation PHS, achieving sophisticated mobile communication to popularize this widely in tourist areas. Another aim is to improve the satisfaction of foreign travelers and promote the tourism industry. In addition to multilingual translation, the project is developing services that are easy for souvenir shops and restaurants in tourist areas to introduce and use, by utilizing such leading-edge technologies as wearable video delivery servers supporting next generation PHS. The project can also take advantage of the fact that it will be used in a special zone to eliminate copyright and other issues in corpus development.

9 Conclusion

Advances in speech and language research have brought speech translation close to the practical level for simple topics where there is a relatively

clear value of use. At the current level, however, speech translation has only reached the stage of creating the core technologies. In order to achieve more sophisticated speech translation, research and development should be further accelerated. Below are some points that should be the subject of focus moving forwards.

Firstly, one distinctive feature of corpus-based technologies is that they improve with use. It is consequently vital to secure opportunities for field and social testing, and to actively use developed technologies. Events like the Olympics and World Fairs, where speakers of many different languages can be expected to attend, are an ideal opportunity to field test multilingual speech-translation technology. It is thus vital to utilize these opportunities to advance the technology. NICT conducted a monitor experiment at the Beijing Olympics, chiefly targeting travelers from Japan. It developed a speech translation system supporting proper nouns in the city of Beijing, and had monitors use speech translation devices to communicate in the city, using the devices for such purposes as transportation, sightseeing, and shopping. A questionnaire was used to survey users' satisfaction with the service, helping to bring speech translation technology a step closer to viability in practical use.

If Japan wishes to become a major tourist destination, then an effective measure could be to provide continuous tourism information services to foreign tourists, using spoken language translation. Meanwhile, the numbers of foreign residents and workers in Japan are increasing, and multilingual speech translation could be an essential means of communication for local governments, medical facilities, police, and education. Even if there are interpreters present, it should help reduce interpreting costs. If all of these systems are used separately, however, then only fragmentary knowledge will be accumulated, making it inefficient in terms of feedback for research and development. It will probably be necessary for national and local government and the private sector to form a cooperative framework in order to improve efficiency. For example, it could be effective to distribute compact translation devices to public institutions where their need is foreseen, and loan these devices to foreign workers and tourists free of charge.

Secondly, speech translation is a technology that translates spoken words in different languages. Although translation into English is of course vital, it will also be highly significant if speech translation can work directly between Japanese and the native languages of many different countries. For this reason, it is vital to increase the numbers of languages supported. There are limits to how far this research and development can progress in Japan alone, especially when it comes to collecting corpora. A scheme of collaboration between countries with many different languages is needed; in other words, a mechanism is needed to enable various countries to work in partnership to research speech translation, speech, and language. Creating international spoken language technology research centers and the like as a scheme for collaborative R&D should render feedback from a wide range of research into the collection of speech and dialect data, language structure, and the like.

Thirdly, when many countries begin to actually research and develop speech translation, it will be necessary to standardize the interfaces to connect these various language processing modules. The development of connection methods, data formats, dictionaries, and the like must maintain an eye toward standardization. We must avoid a situation in which each country develops its own system, and the systems are not mutually compatible. Speech translation technology is advanced in Japan, and this country can thus lead other countries with relation to standardization.

Finally, attention must be given to copyright. Speech and language processing require speech and text corpora, and the performance of speech translation depends heavily on the quantity and quality of these corpora. Consequently, the use of corpora of news broadcasts, newspapers, and the Internet is extremely effective. Current copyright law does not take secondary uses such as these types of corpora into account. In order to research and develop new technologies, it will be necessary to revise and administer the law so that it is more flexible. This topic is currently being debated by the Copyright Working Group of the Cultural Council, and a conclusion will be published in the near future. It will be necessary to systematically reorganize the topics for future full-scale speech translation services, and reconsider the response, including service models, after the results of this study are released.

References

[1] S.Nakamura et al,"ATR Multi-lingual Speech-To-Speech Translation System",IEEE Trans. ASLP, vol.14, no. 2 (2006)

[2] A. Finch et al, "The NICT/ATR Speech Translation System for IWSLT 2007", IWSLT (2007)

[3] T. Takezawa et al, "Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World" Proc. LREC (2002)

[4] Yasuda et al,"Automatic Translation Evaluation Technologies and Their Application for the Research and Development of Automatic Translation", Journal of Japanese Society for Artificial Intelligence, Vol. 23, Issue 1 (2008) (Japanese)

[5] Sugenoya et al,"Proposal for a Method to Evaluate the Speech Translation Capability of Speech Translation Systems and Humans, and Comparative Experiment", IEICE Transactions J84-D-II, 11 (2001) (Japanese)

[6] Itoh et al, "Field Study and Evaluation of Japanese-English-Chinese Speech Translation Device", 1-Q-33, Proceedings of the Acoustical Society of Japan Spring Meeting (2008) (Japanese)

[7] IWSLT : http://www.slt.atr.jp/IWSLT2004/

[8] SLTC e-Newsletter,"DARPA's GALE Program to Get More Challenging in 2007" (Example from GALE Project) :
http://ewh.ieee.org/soc/sps/stc/News/NL0701/NL0701-GALE.htm

[9] ETSIES 202 050 ETSI ES 202 050 v1.1.1 Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI, April 2002.

[10] APEC TEL WORKING GROUP :
http://www.apectelwg.org/

[11] ASTAP :
http://www.aptsec.org/Program/ASTAP/

[12] Kimura et al,"Considerations for a Communication Interface for a Multilingual Speech Translation Platform", 3-Q-17, Proceedings of the Acoustical Society of Japan Autumn Meeting (2007) (Japanese)

Profile

Satoshi NAKAMURA
Affiliated Fellow,NISTEP

Executive Researcher, National Institute of Information and Communications Technology, Japan
Project Leader, MASTAR Project
Director, ATR Spoken Language Communication Research Laboratories
ATR Fellow
http://www2.nict.go.jp/x/x162/mastar
Doctor of Engineering. Currently involved in speech and language processing, with a focus on speech translation and speech recognition. Visiting Professor at University of Karlsruhe, Germany and Professor, Keihanna Joint Graduate School.

(Original Japanese version: published in August 2008)