Multilingual Speech Translation

Shoichi Senda
Technical Expert
Standardization Promotion Office, Innovation Promotion Department
National Institute of Information and Communication Technology

1. Introduction

Speech communication is the most basic means of communication used by humanity, and telephone services implementing speech communication have been steadily enhanced to the point where it is available "anywhere, anytime, and with anyone." This progress has overcome almost all technical issues related to "anywhere" and "anytime", but implementing "with anyone" involves language differences, which are a major obstacle, and we have not yet reached the point where we can communicate freely with anyone using speech. Actually implementing an automatic translation telephone has been a common human dream since the telephone was invented. Anyone would want such a technology and there have been various attempts and much research in this field, but only now are we beginning to see that it could become practical soon.

With the opening of the Tokyo Olympics and Paralympics, ever increasing numbers of foreign visitors are expected and there is hope that, improving multilingual speech translation technology to a sufficiently practical level, promoting it, and implementing it in society, will facilitate communication with visitors and enable them experience Japanese hospitality directly.

2. History and standardization in the study of speech translation

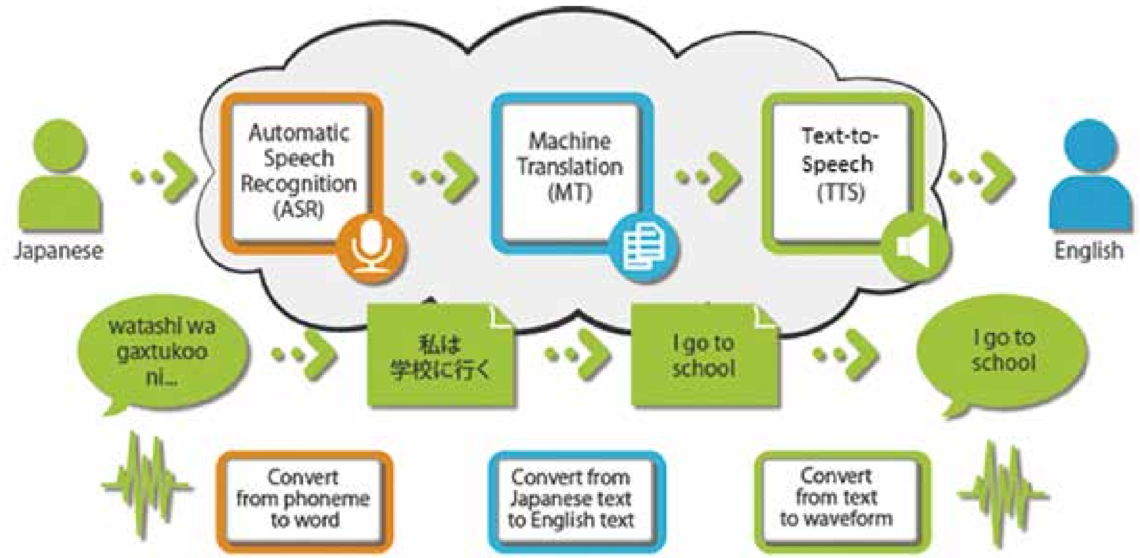
Here we look at the history over many years, of R&D attempting to realize the dream of multilingual speech translation, and also the part played by standardization.

At the beginning, different research agencies studied speech translation separately, but individual study had limitations, so in 1991, scientists from around the world, including the USA, Germany, France, Japan, Korea, and Italy, began forming voluntary organizations such as the Consortium for Speech Translation Advanced Research (C-STAR), and they started collaborating by sharing independent research results for each language and integrating them together. Thus, speech translation R&D started in many countries, and an awareness of the need for standard interfaces and data formats grew, to ensure intercompatibility of their work. In particular, the Asian Speech Translation Advanced Research Consortium (A-STAR) was organized in 2006, centered on national research agencies from six Asian countries having many official languages within their regions. The Asia-Pacific Telecommunity (APT) Standardization Program (ASTAP) also began standardization activities. Before long, it became clear that this standardization activity at ASTAP should be done globally, and not be limited to the Asia-Pacific region, so activity was moved to ITU-T SG16, and standardization on a global scale began. As a result, in 2010, recommendations ITU-T F.745, specifying functional requirements for network based Speech-to-Speech Translation (S2ST), and H.625, specifying architectural requirements, were created. A-STAR, which had been restricted to the Asia-Pacific region, was also expanded and reorganized to continue its activities as the Universal Speech Translation Advanced Research consortium (U-STAR). Recently, with advances in Big Data analysis and AI, this once-limited field has begun to produce practical products and services.

3. Implementation of multilingual speech translation

Generally, to realize two-way speech communication between people speaking different languages, the audio signal expressed in the speaker's language must be translated to an audio signal in the language of the listener. For example, Japanese to English translation is implemented as in Figure 1. When the speaker says "watashi wa gaxtukooni…" in Japanese, the audio signal is automatically recognized as "わたしは学校に行く" in Japanese text, this text is machine translated to "I go to school" in English that the listener can understand, and the text is then converted to an English audio signal.

Figure 1: Outline of speech translation


Japanese
Automatic Speech Recognition (ASR)
Machine Translation (MT)
Text-to-Speech (TTS)
English
watashi wa gaxtukoo ni...
私は学校に行く
I go to school
I go to school
Convert from phoneme to word
Convert from Japanese text to English text
Convert from text to waveform

Here, we have described speech translation as implemented by performing Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) functions in sequence, using the case of Japanese to English translation as an example. This combination of functions can be used to translate between any languages and is not limited to Japanese-to-English. Here, ASR converts the input speech signal into text data, MT converts the input text data into text data in a different language (but having equivalent meaning), and TTS converts the input text data into a speech signal. The details of these conversions depend on the languages and fields being translated. As such, ASR, MT and TTS generally share the basic function of using conversion dictionaries called corpuses for each translated language and field. When building a multi-lingual speech translation system in this way, the engines that realize each of these functions can be designed and developed independently, and multi-lingual speech translation can be achieved more easily by combining engines appropriately, according to the languages and application area (Figure 2).

Figure 2: Architecture of multilingual speech translation

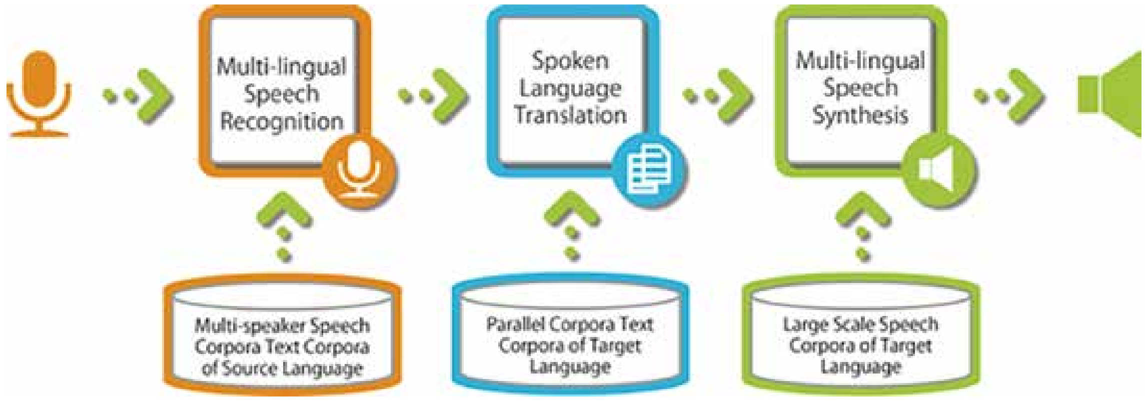
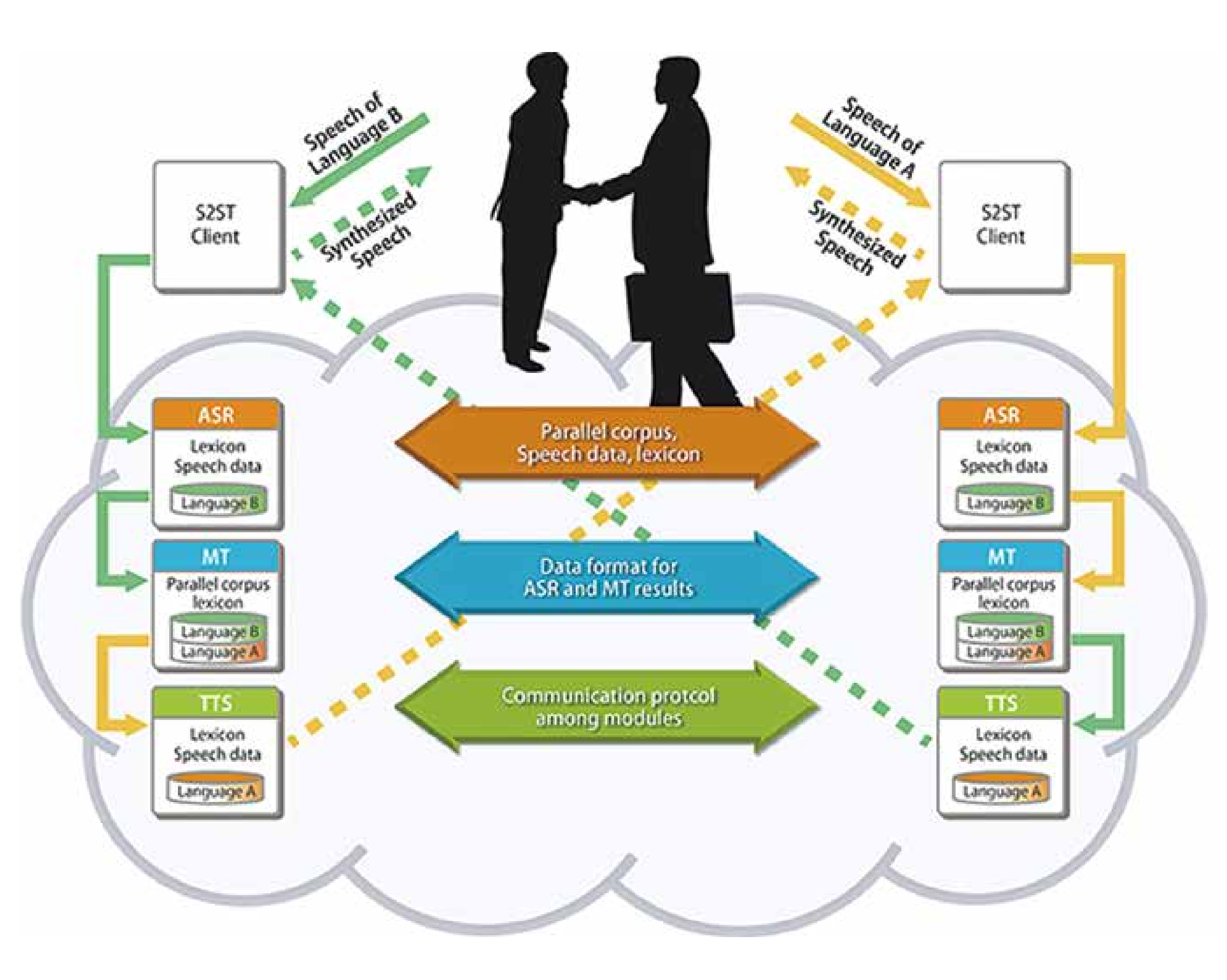

Multi-lingual Speech Recognition
Spoken Language Translation
Multi-lingual Speech Synthesis
Multi-speaker Speech Corpora Text Corpora of Source Language
Parallel Corpora Text Corpora of Target Language
Large Scale Speech Corpora of Target Language

Figure 3: Architecture and its implementation protocol



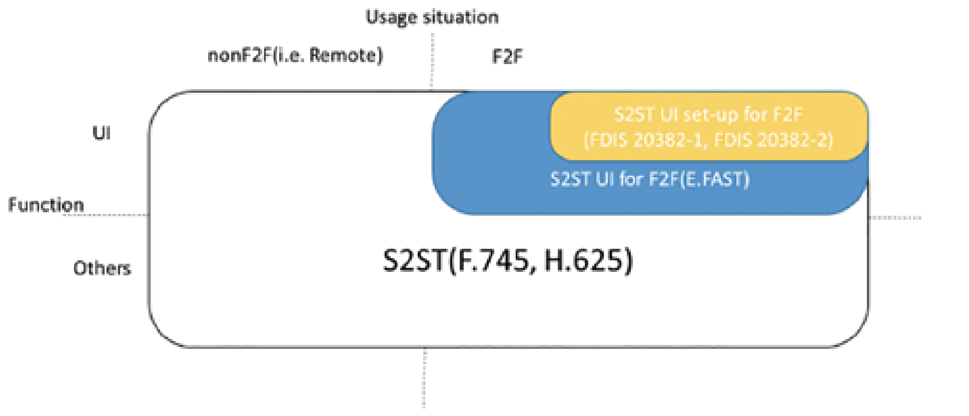
To make it possible to implement multilingual speech translation with this sort of structure, the functionality of each functional element be standardized, as well as how each functional element is deployed, what and how they exchange information, including protocols and formats. Since the main functional elements of speech translation, namely ASR, MC, and TTS, require significant amounts of processing, the ITU-T SG16 has adopted a client-server model in which servers are placed on the network and users use server functionality through clients that are installed on their various terminals. As such, protocols between client and server, and between servers, are being standardized (Figure 3).

4. State of standardization in speech translation

Later in the speech translation field, in 2014, ITU-T SG2 Q4 and ISO/IEC JTC1/SC35 WG5 proposed that user interfaces be studied and standardized to improve service availability, independent of system functional and organizational requirements. These standardization proposals regulate user interfaces particularly for when speaker and listener are engaged in two-way communication with each other at a single location. S2ST has already been standardized and has moved to a maintenance and management phase, so the relation of these proposals to the completed recommendations, F.745 and H.625, as well as any particular requirements due to the face-to-face environment needed to be clarified. In this area, there were also proposals regarding the user interface for speech translation in face-to-face environments from ASTAP EG-SNLP in Q4/2 (which was later integrated into EG-MA), which was studying the E.FAST draft recommendation. It was agreed that more investigation was needed to understand various use cases and associated issues, and that work should move from SG2 to SG16, become Q24/16, and continue in collaboration with Q21. On the other hand, speech translation user interfaces being studied in ISO/IEC JTC1/SC35 WG5, IS20382-1 and -2 were standardized without necessarily coordinating sufficiently with ITU-T speech translation standardization activities, and ITU-T submitted comments identifying issues during DIS voting. The reply given to ITU-T SG16 by JTC1/SC35 was that these standards apply only to the user interface setup. In response, ITU-T SG16 reviewed and reorganized the interrelation among S2ST related standards (F.745, H.625, E.FAST, IS 20382-1 and -2) as shown in Figure 4, and intends to reaffirm this at the JTCI/SC35 meeting in February, 2018.

Note that while F.745 and H.625 were standardized with the goal of realizing two-way speech translation, input and output signals are not necessarily limited to audio signals. By skipping the ASR and TTS functions of F.745 and H.625, and using only the MT function directly for input and output text data, these standards could be used for multilingual text translation systems. Or, by skipping the ASR function for communication in one direction, and the TTS function in the other direction, tools such as Koetra (http://www.koetra.jp/) can be created to support communication between deaf and hearing people. Still further, by applying recent advances in Big Data, AI and Deep Learning to video recognition, the ASR function could be replaced or extended, from simple speech recognition to recognition of sign-language video, so that multi-lingual speech translation could evolve into general translation, including sign language. Even if this will be difficult to realize by the time the Paralympics begin, continuing to work on implementing this sort of tool in support of disabled persons has the benefit of contributing to the hopes and dreams of many who live with such disabilities.

Figure 4: Interrelation among S2ST related standards



F.745 and H.625 (ITU-T Q21/16) cover both of remote and F2F situation and do not exclude UI function.
E.FAST (ITU-T Q24/16) covers UI function in F2F situation
IS 20382 (JTC1/SC35) covers F2F UI setup among two or more S2ST systems