

APT REPORT

on

Harmonization of S2ST (Speech-to-Speech Translation) Standardization

**APT/ASTAP/REPT-40**Edition: June 2019

**Adopted by**

**31st APT Standardization Program Forum (ASTAP-31)**

**11 - 15 June 2019, Tokyo, Japan**

(*Source: ASTAP-31/OUT-17*)

**Contents**

[**1. Introduction** 3](#_Toc11936397)

[**2. History and standardization in the study of speech translation** 3](#_Toc11936398)

[**3. Implementation of multilingual speech translation** 5](#_Toc11936399)

[**4. State of standardization in speech translation** 6](#_Toc11936400)

[**5. Future activities** 7](#_Toc11936401)

[List of abbreviation 7](#_Toc11936402)

[References 7](#_Toc11936403)

[Appendix A: Use cases 9](#_Toc11936404)

[A1. Use case in Japan: Service of “OMOTENASHI GUIDE” 9](#_Toc11936405)

[A2. Introduction of “Speech translator/simultaneous interpretation service” 10](#_Toc11936406)

[Appendix B: Issues on Use cases of speech translation 12](#_Toc11936407)

**1. Introduction**

Verbal communication is the most basic human communication, and telephone service have been making verbal communication possible “anywhere, anytime, and with anyone” However, communication “with anyone” sometimes involves language barriers. Implementing an automatic translation has been a common human dream since telephone was invented.

**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 in 1991, scientists from around the world began forming voluntary organizations such as the Consortium for Speech Translation Advanced Research (C-STAR), and they started collaboration to integrate research results of each language. Then, there emerged the need for standard interfaces and data formats which ensure compatibilities 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 who have many official languages within their regions.

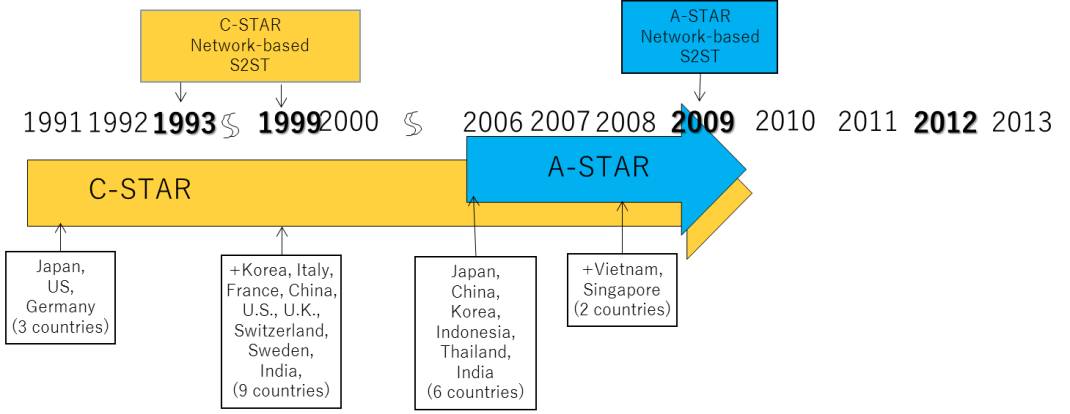


Fig.1: Initiation of International Consortium

The Asia-Pacific Telecommunity (APT) Standardization Program (ASTAP) also began standardization activities. Then it became clear that this standardization activity should be globalized rather than limited to the Asia-Pacific region. The activity was moved to ITU-T SG16, to make it global. 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).



Fig. 2: Standardization from Asia

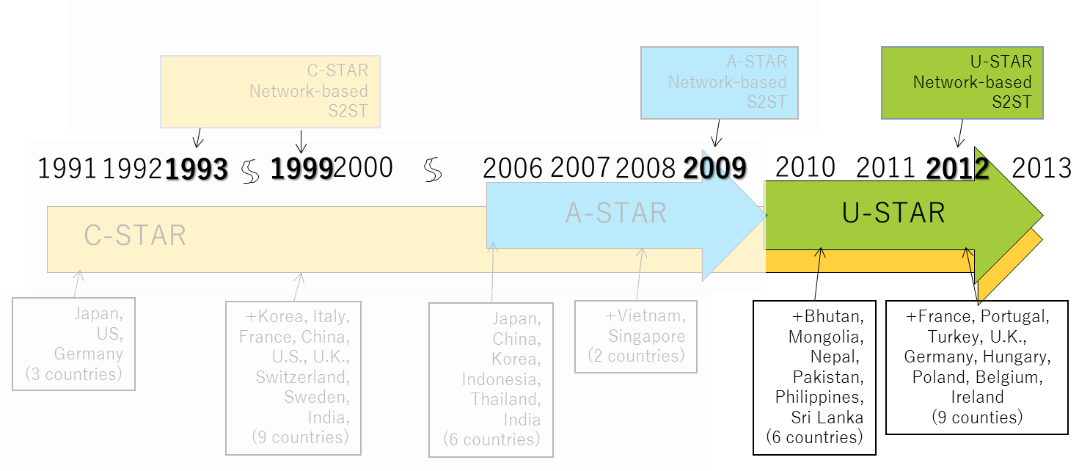


Fig.3: Transformation from A-STAR to U-STAR

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. Recently, with advances in Big Data analysis and AI, this once-limited field has begun to produce practical products and services.

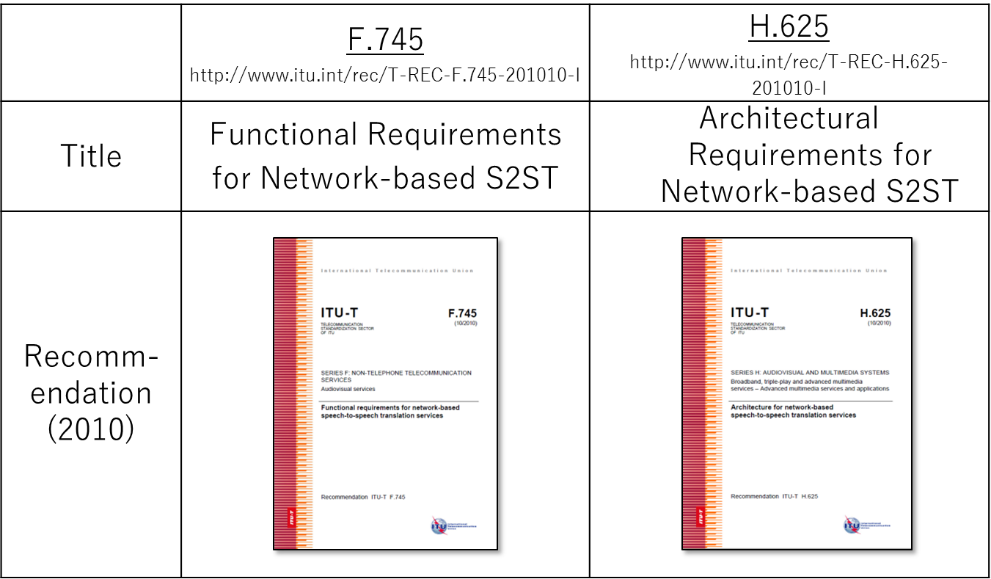


Fig.4: ITU-T Standards

**3. Implementation of multilingual speech translation**

Generally, to realize automatic two-way verbal 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 below. When the speaker says “watashi wa gaxtukooni…” in Japanese, the audio signal is automatically recognized it as “私は学校に行く” 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.

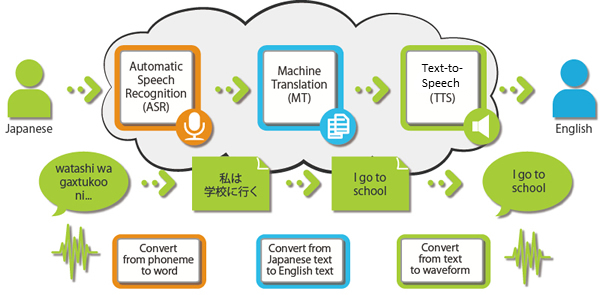


Fig.5: Overview of speech translation

Here, we have described speech translation composed of 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 applied to translation between any two 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 areas.

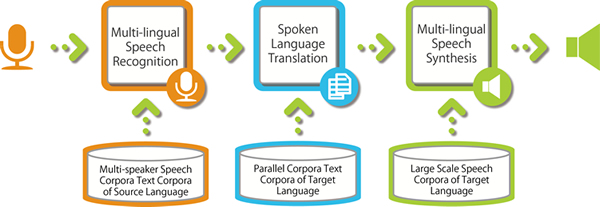


Fig. 6: Multi-lingual speech translation system architecture

To implement multilingual speech translation, the each functional element should be standardized, as well as how each element is deployed, how they exchange what 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.

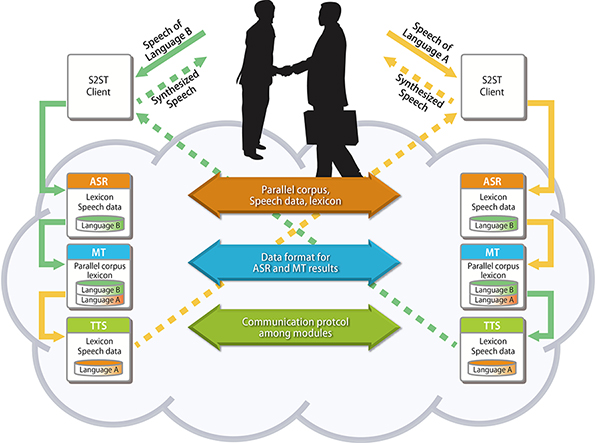


Fig. 7: Architecture and its supporting protocols

**4. State of standardization in speech translation**

In 2014, ITU-T Q4/SG2 and ISO/IEC JTC1/SC35 WG5 respectively proposed to study and standardize user interfaces to improve service availability, independent of system’s functional and organizational requirements. These standardization proposals regulate user interfaces particularly when a speaker and a listener are engaged in two-way communication with each other at the same place. ITU-T SG16 reconfirmed the interrelation among S2ST related standards (F.745, H.625, E.FAST, IS 20382-1 and -2) as shown in figure below.

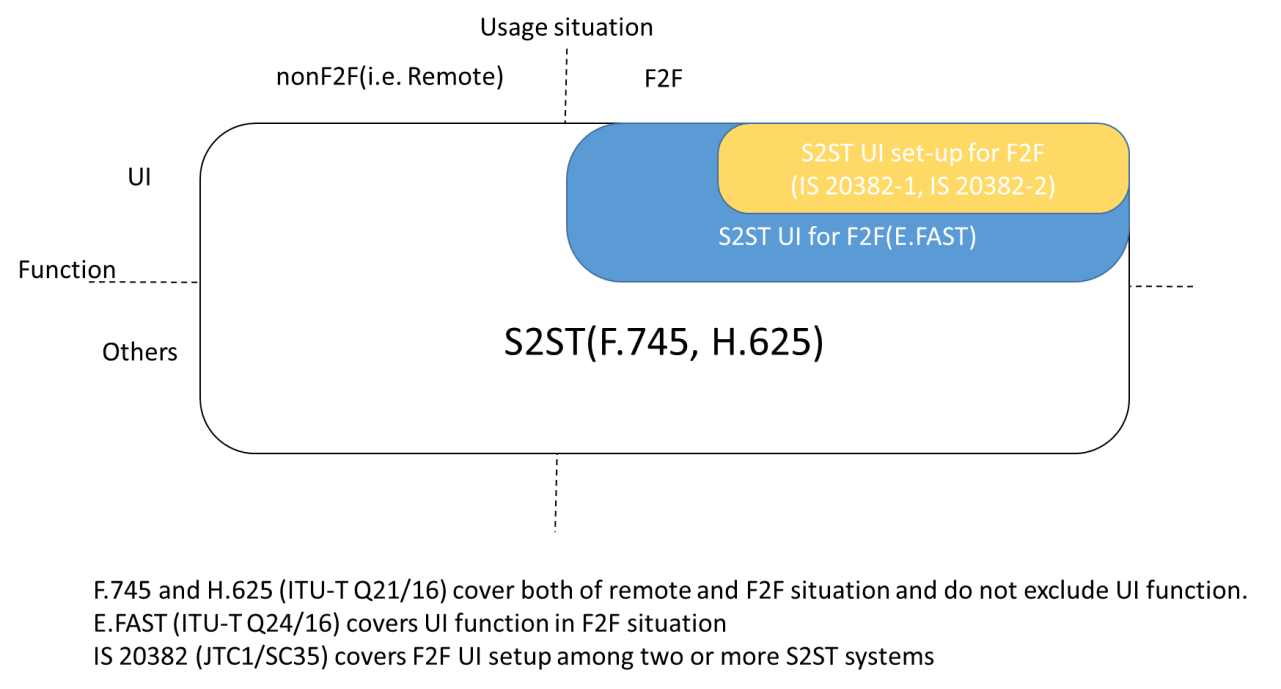


Fig.8: interrelation among S2ST related standards

**5. Future activities**

Though F.745 and H.625 were targeting 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/en/) 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.

# List of abbreviation

AI Artificial Intelligence

ASR Automatic Speech Recognition

A-STAR Asian Speech Translation Advanced Research Consortium

C-STAR Consortium for Speech Translation Advanced Research

MT Machine Translation

S2ST Speech-to-Speech Translation

TTS Text-to-Speech

U-STAR Universal Speech Translation Advanced Research Consortium

# References

[ITU-T F.745] Recommendation ITU-T F.745 (2010, 2016), *Functional requirements for network-based speech-to-speech translation services*.

[ITU-T H.625] Recommendation ITU-T H.625 (2010, 2017), *Architecture for network-based speech-to-speech translation services*.

[ITU-T E.FAST] Draft Recommendation ITU-T E.FAST (under study), *Architecture for network-based speech-to-speech translation services*.

[ISO/IEC 20382-1] ISO/IEC 20382-1 (2017), Information technology – User interfaces – Face-to-face speech translation– *Part 1: User interface*.

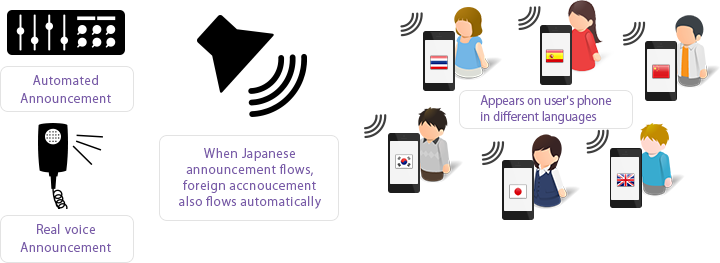
[ISO/IEC 20382-2] ISO/IEC 20382-2 (2017), Information technology – User interfaces – Face-to-face speech translation– *Part 2: System architecture and functional components*.

**Appendix A Use cases**

1. Use case in Japan: Service of “OMOTENASHI GUIDE”

A1.1 Description of the service

Omotenashi Guide provided by YAMAHA (Yamaha Corporation) is a system for supporting a “universal design of sounds”, with which everyone independent from his/her nationality can understand Japanese language. In facilities where Omotenashi Guide is implemented, any voice and text information in multi-language form can be provided on screens of user’s smartphones by simply making Japanese announcements. Since relevant information is transferred in the form of sounds, it is possible to send translated voice and text information to users’ smartphones even if the Internet access is not available. Figure A-1 shows an illustrative explanation of Omotenashi Guide.

****

Source: YAMAHA homepage, http://omotenashiguide.jp/en/

Figure A-1. Illustrative explanation of Omotenashi Guide

The current main concern in Japan is how effectively Japanese voice information such as public announcements is transferred. When people think about the improvement of convenience and/or user-friendliness for visiting foreigners and the realization of society where elderly people and/or people with hearing disability can live comfortably, the effective transmission of Japanese voice information would be one of the most urgent issues that have to be dealt with. To this end, one might say that multiple announcements with different languages could be announced, or voice information could be displayed on large LCD panels and sign boards. However, such solutions would not be realistic from the cost perspective or limited conditions with time and space.

NICT and YAMAHA have been jointly developing the system since July, 2015. Under this joint development, the goal would be to develop functions such as:

* Contents in Japanese announcement can be automatically recognized simply by speaking to microphones
* Voice and/or text information with appropriate foreign language can be automatically selected from an announcement book.

A1.2 Languages supported

Japanese, Chinese, English, and Korean languages are currently supported.

A1.3 Technologies used in the service

Omotenashi Guide is based on the following technologies:

* Voice recognition of public announcement from speakers installed in buildings such as airports and shopping centers, and making it text
* Playing the announcement from speakers after converting the text into sounds whose frequency range is 18 – 20 KHz, which is outside a human audible band, then
* Displaying the text after collecting the sounds by a microphone of any smartphones

In addition, the following technologies are utilized as well.

* Automatic translation of real time announcements that are not provided as programed broadcastings
* Displaying its contents on a smartphone’s screen

A1.4 Deployment examples

Omotenashi Guide is available as a field trial service in such areas as airports, trains, busses, shopping centers, and tourist spots. It is also noted that Omotenashi Guide is officially adopted at Japan Pavilion in Expo 2015 Milano, which has opened since 1st May, 2015.

A1.5 References

http://www.nict.go.jp/press/2015/07/13-1.html

http://omotenashiguide.jp/

1. Introduction of “Speech translator/simultaneous interpretation service”

A2.1 Service overview

Toshiba corporation provides interactive voice translation service “RECAIUS speech translator/simultaneous interpretation service” by utilizing original cloud-based AI service for audio and visual called RECAIUS. This system can be used in many scenes such as information services on goods and venue, and customer care in events, tours, and shops, in order to support communication with visitors from foreign countries. This is also useful in public transportation services and commercial facilities.

The “RECAIUS speech translator” is an application for smart phones. Talked speeches to a smart phone are translated to other languages, the translated texts in other language are displayed, and the translated speeches can be listened from the smart phone. Different from ordinal translation applications for tourists, the feature of this application is to help staffs in shops and venues where many foreign visitors come to. When staffs and visitors speak to a smart phone alternately, the content is automatically translated in text and voice in real-time. This allows them to have conversations while viewing translation results on the smartphone screen (See Fig.1). Its main functions include dictionary registration and phrase registration

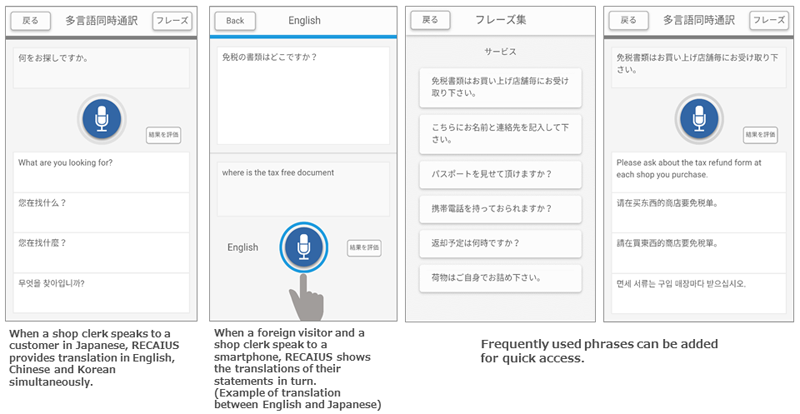


Figure 1 Display images in RECAIUS speech translator

The “RECAIUS simultaneous interpretation service” provides speech recognition for speech translation, colloquial translation, and voice synthesis engine through Web API. By concurrent processing of speech recognition and machine translation, the feature is to translate speeches in real-time. The accuracy of translation can be improved by customizing dictionary for intended purpose on specific scenes and businesses. Translated results can be displayed by texts and also answered by synthetic voices.

A2.2 Use cases

From 2014, Toshiba has worked positively to field test of above services together with railway companies and shops. The field test aimed to verify accuracy of guidance for foreign tourists, by installing tablet terminals used for RECAIUS simultaneous interpretation services into train stations and shops, and by travelling staffs carried the tablet terminals. On January 2016, commercial services of simultaneous interpretation, which were provided to smart phones, were launched for information counters in shopping centers.

In order to improve quality of “omotenashi” services, meant “hospitality” services, on translation/interpretation services from the viewpoint of foreign tourists, actual data on communication at the real fields will be analyzed. For example, analyzing the stored data might clarify what problems for tourists from which countries are, and what major inquiries from tourists are. Based on the results, continuous improvement, such as utilizing digital signage system for clear guidance on information which might make confusion for foreigners, aimed to provide better services to foreign visitors to Japan.

A2.3 Major features

The following functions are provided from cloud system, RECAIUS:

* Simultaneous interpretation function on multi-languages

Japanese can be interpreted to English, Chinese and Korean at the same time, considering the user could not understand visitors’ language. It is possible to provide the results on voice.

* Phrase registration function

In order to show translation results rapidly, commonly used phrases can be registered to the system.

* Speech recognition function

It is possible to use speech recognition by WebAPI (Rest API). The following formats are supported:

* Sound format: 4bit ADPCM, 16bit liner PCM, and Speex
* Sampling frequency: 16kHz
* No of channels: 1 (monaural)
* Colloquial translation function

By using consecutive translation, it is possible to have natural conversation even if long or complicated sentences are spoken.

* Voice synthesis function

Voice synthesis can be provided by WebAPI (Rest API). The output formats supported are:

* WAV
* OGG
* M4A
* Dictionary registration function

Since staffs in shops can register commonly used words into the dictionary of the system, the names of goods and facility can be translated correctly.

A2.4 Languages supported

* Speech recognition function: Japanese, English and Chinese(Mandarin)
* Interpretation function: Japanese 🡨🡪 English and Japanese 🡨🡪 Chinese,   
  (Japanese 🡨🡪Korean plans to be provided in 2017)
* Speech synthesis function: Japanese, English, Chinese(Mandarin) and Korean

A2.5 Operating environment

* RECAIUS speech translator

OS: Android（5.0, 5.1, 6.0, 7.0）and iOS（8.4, 9.1, 9.2, 9.3, 10.0, 10.1）

Testing terminal： Nexus6（AndroidOS5.1.1）

* RECAIUS simultaneous interpretation service

Proving WebAPI

\*recommended web browser for service maintenance site: Internet Explorer 11, the latest version of Chrome, Firefox 35 or above, Safari 8.0.3

A2.6 References

http://www.toshiba.co.jp/cl/en/news/news20161019.htm

http://www.toshiba.co.jp/cl/pro/recaius/case.html

http://www.toshiba.co.jp/cl/pro/recaius/lineup/interpreting.html

http://www.toshiba.co.jp/cl/pro/recaius/lineup/creator.html

http://www.toshiba.co.jp/cl/news/news201510\_02.htm

http://www.toshiba.co.jp/iot/power/entry/2016/2016\_016\_01.htm

http://www.toshiba.co.jp/iot/power/entry/2016/2016\_016\_02.htm

**Appendix B Issues on Use cases of speech translation**

This appendix indicates some speech translation use cases which still has some issues to be considered.

The typical Japanese use cases of SNLP technologies, especially speech-to-speech translation systems are uses for announcement at transportation facility, and for counter services in shopping and medical scenes. The features differ from individual use cases, and issues on user interfaces (UIs) to be considered also differ. Use cases can be divided into two situations, and one-to-many situations can also divided by user interface types, the types are shown as follows:

* + one-to-one situation: this includes conversation and interview
  + one-to-many situation: this includes announcement and guidance
    - one-to-many situation with individual user interface: the user may use own terminal
    - one-to-many situation with a single common user interface: all users commonly use a single human interface simultaneously.

Table B-1 summarizes issues to be considered for each typical use case, and one of them was introduced at previous ASTAP meetings.

Table B-1 Typical use cases of multi-language speech translation and issues to be considered

|  |  |  |  |
| --- | --- | --- | --- |
| **Use cases** | **Overview** | **Major feature** | **Issues** |
| Counter services in shopping centers  (for information of goods and venue) | **[One-to-One situation]**  Conversation with translated voice and text on tablet terminals or smart phones at the information counters (visitors can communicate on speech and text with staffs) [1]. | * Functions of speech recognition, colloquial translation and voice synthesis engine, which are required for speech to speech translation, can be provided by Web API. * Application recognizes language type and displays spoken and translated texts automatically, when visitors ask something to staffs in selected languages. | * Languages of visitors are unknown for counter parson in advance. * Staffs may not be able to speak visitors’ languages. * Visitors may not be familiar for goods and venue. * There might be existed frequent asked questions. * Staffs are allocated in whole shopping center, not only in counter. |
| Counter services in hospitals  (Registration desk, medical examination, hospital ward, accounting, and etc.) | **[One-to-One situation]**  Conversation via tablet terminals without any operation, which have capability of hands-free speech to speech translation, in several scenes in hospital. | * The system requires not for operating the terminal, especially for preventing communicable disease. | * The system is required to provide several ways for output translation results. * The system is required to be used even in noisy environment. * It is needed to prevent communicable disease by touching user terminals. |
| Counter service in municipal offices | **[One-to-One situation]**  A speech-to-speech translation app specifically designed for municipal offices to meet the growing demands due to the increase of the number of foreign nationals in Japan.  The app can be installed on iOS/Android smart devices and enables smooth communication between the Japanese office staff and foreign nationals.  [2] | * The speech-to-speech translation app "Voice Biz" supports more than 30 languages (11 of which supports speech input) to be translated from/into Japanese. The app also contains a set of frequently-used phrases and proper nouns which can be customized according to the purpose of use. Not only is the app time efficient, but is capable of providing accurate information about the area where the app is being used. | * Further promotion is planned as the app aims to be to be widely adopted in municipal offices throughout the country. |
| Emergency measures (first-aid treatment for foreign nationals and tourists in Japan) | **[One-to-One situation]**  A translation app specifically designed for emergency medical teams in Japan to support communication between foreign patients when providing first-aid treatment.  The app can be installed on iOS/Android smart devices and enables smooth communication between Japanese emergency staff and foreign patients during first-aid treatment. "First-aid treatment" refers to the point from when the staff arrive on site after receiving the emergency call, up until they take the patient to the appropriate hospital with an ambulance.  Since its release in April 2017, the app is being used in more than half (54%) of the 726 emergency medical teams (fire departments) in Japan (as of April 2019).  [3] | * The app contains a set of pre-registered fixed phrases which can be translated into any of the 15(\*) supported languages in an instant without network connectivity. * The phrases are basically yes/no questions where the staff can select and show the patient (or attendant, in cases where the patient is unconscious or badly injured) in the designated language. * (e.g., Q: "What were you (the patient) doing when the symptoms started? / A: 1) Sleeping, 2) Eating, 3) Other ) * This operation method is effective not only when collecting necessary information with not a moment to spare, but for cases when informed consent is required to provide certain medical treatments. The app is also capable of speech-to-speech translation which is useful when the patient needs to communicate with the emergency staff about matters that are not covered by the fixed phrases.   (\*) English, Mandarin, Korean, Spanish, French, Thai, Indonesian, Vietnamese, Myanmar, Taiwanese Mandarin, Malay, Russian, German, Nepali, and Brazilian Portuguese. | * Issues concerning noisy environments: * While the translation results are also displayed in text, audio (synthesized speech) is sometimes difficult to hear in noisy environments. Speech recognition errors (when using speech input) may also occur in similar situations. |
| Announcement (in airports, in train stations, inside buses) and Guidance  (at events sites) | **[One-to-Many with individual UI]**  Displaying translated texts, which are prepared in advance, on users’ smart phones by using existing communication or announcement facility of the venue [4]. | * The system may easily be installed by utilizing existing communication or announcement facility. * Users who use foreign languages can immediately get information in their languages by text messages on the application of smart phones. These messages might be transmitted by audio frequency over announcement facilities. | * It is required that users can installed the application easily as much as possible. * Information should be provided visitors’ favorite languages at least in text message. * It is desirable that users can customize or set configuration of the application by themselves. |
| **[One-to-Many with a single common UI]**  Guiding visitors by using megaphone-type speech to speech translator [5]. | * Considering situation, it can be used even in emergency situation. | * Since one announcement is repeated in several languages, foreign visitors might not get information immediately. |

As shown in the Table 1, there are different issues for each situation, so the requirements for implementing speech-to-speech translation in face to face situations are need to be considered for each situation. The requirements of the user interfaces should be considered the followings, but not limited to:

* Whether a single common interface shared by participants or individual interfaces for each participant are used?
* If individual interface are used, does it have availability of customization for each interface?
* Whether simultaneous interpretation for multiple languages or consecutive translation sequentially for several languages is required for users?
* Whether a movable interface with simple functionalities or a fixed interface with rich functionalities is used?

Especially, in medical scene, requirements shall be strictly considered since this might be related to human life. For example, medical interview in hospital as an example of one-to-one situation makes medical decision for patients by doctors, so the translation system for medical scene should be considered for noisy environment and to provide several ways for output translation results. In addition, the system is recommended to be operated by hands-free, in other wards touch-less operation, because medical staffs including doctors usually use their both hands for providing medical care, and also because hands-free operation might prevent communicable disease both for patients and medical staffs.

References for appendix B

[1] http://www.toshiba.co.jp/cl/en/news/news20161019.htm

[2] <https://www.toppan.co.jp/solution/service/VoiceBiz.html>

[3] http://www.soumu.go.jp/main\_content/000553523.pdf

[4] ASTAP-27-INF-04, NTT-AT, “INTRODUCTION OF A USE CASE IN JAPAN FOR SPEECH AND NATUAL LANGUAGE PROCESSING TECHNOLOGIES”

[5] http://news.panasonic.com/global/topics/2016/45751.html