International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

F.745

(07/2016)

SERIES F: NON-TELEPHONE TELECOMMUNICATION SERVICES

Multimedia services

Functional requirements for network-based speech-to-speech translation services

Recommendation ITU-T F.745

ITU-T

International Telecommunication Union

ITU-T F-SERIES RECOMMENDATIONS

NON-TELEPHONE TELECOMMUNICATION SERVICES

TELEGRAPH SERVICE	
Operating methods for the international public telegram service	F.1–F.19
The gentex network	F.20–F.29
Message switching	F.30–F.39
The international telemessage service	F.40–F.58
The international telex service	F.59–F.89
Statistics and publications on international telegraph services	F.90–F.99
Scheduled and leased communication services	F.100–F.104
Phototelegraph service	F.105–F.109
MOBILE SERVICE	
Mobile services and multidestination satellite services	F.110–F.159
TELEMATIC SERVICES	
Public facsimile service	F.160–F.199
Teletex service	F.200–F.299
Videotex service	F.300–F.349
General provisions for telematic services	F.350–F.399
MESSAGE HANDLING SERVICES	F.400–F.499
DIRECTORY SERVICES	F.500–F.549
DOCUMENT COMMUNICATION	
Document communication	F.550–F.579
Programming communication interfaces	F.580–F.599
DATA TRANSMISSION SERVICES	F.600–F.699
MULTIMEDIA SERVICES	**F.700–F.799**
ISDN SERVICES	F.800–F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850–F.899
HUMAN FACTORS	F.900–F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.745

Functional requirements for network-based speech-to-speech translation services

Summary

Recommendation ITU-T F.745 specifies a high level functional model, a service description and requirements for speech-to-speech translation (S2ST) accomplished by connecting distributed S2ST modules all over the world through a network. To extend this network-based S2ST to other modalities, such as sign language, the modality conversion markup language (MCML) needs to have an expandable structure. The scope of this Recommendation is limited to the application protocol and the services using the network-based S2ST.

This revision includes additional information to clarify that Recommendation ITU-T F.745 could be applicable to both face-to-face communication and remote communication.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T F.745	2010-10-14	16	11.1002/1000/10982
2.0	ITU-T F.745	2016-07-14	16	11.1002/1000/12897

Keywords

Automatic speech recognition (ASR), machine translation (MT), modality conversion markup language (MCML), speech-to-speech translation (S2ST), text-to-speech synthesis (TTS).

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Introduction

The fact that the world has many different languages is one of the barriers to mutual understanding. The more directly people who speak different languages can communicate without language boundaries, the more mutual understanding can be accelerated and the closer human relationships can be constructed all over the world. To achieve such communication between humans, speech-to-speech translation (S2ST) technologies can be used. S2ST is a technology that recognizes the speech in one language, translates the recognized speech into another language, and then synthesizes the translation into speech. The leveraging of S2ST technologies in a pragmatic manner, which has long been one of mankind's dreams, may have a significant impact on tourism, social services, safety, and security by removing language barriers, and may ultimately influence language education. To construct S2ST systems, automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) must be built for source and target languages by collecting speech and language data, such as audio data, its manual transcriptions, pronunciation lexica for each word, parallel corpora for translation and so on. It is very difficult for individual organizations to build S2ST systems covering all topics and languages. However, by interconnecting ASR, MT and TTS modules developed by separate organizations and distributed globally through a network, one can create S2ST systems that break the world's language barriers.

This Recommendation defines the service description and the requirements for network-based S2ST technologies consisting of various distributed modules connected together in a network.

Recommendation ITU-T F.745

Functional requirements for network-based speech-to-speech translation services

1 Scope

This Recommendation specifies the service description and the requirements for speech-to-speech translation (S2ST) accomplished by connecting distributed S2ST modules all over the world through a network. This service provides S2ST that recognizes the speech in one language, translates the recognized speech into another language, and then synthesizes the translation into speech. People who speak different languages can communicate using this service.

The applications and services using network-based S2ST technologies are characterized by the following components:

– S2ST client:
  - user client for speech/text input and output.

– S2ST servers:
  - speech recognition: speech is recognized and transcribed;
  - machine translation: text in source language is translated into text in target language;
  - speech synthesis: speech signal is created from text.

– Communication protocol:
  - communication protocol to connect user clients and the above S2ST servers.

In order to extend the network-based S2ST to other modalities (e.g., sign language), a communication protocol is incorporated for modality conversion (MC), which converts single/multiple modality information to different single/multiple modality information. The communication protocol for MC needs to have an expandable structure.

– Modality conversion markup language (MCML):
  - XML schema that serves as a data description for data exchanged among modality conversion modules.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

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

[IETF RFC 2279] IETF RFC 2279 (1998), *UTF-8, a transformation format of ISO 10646*.

[IETF RFC 2396] IETF RFC 2396 (1998), *Uniform Resource Identifiers (URI): Generic Syntax*.

[IETF RFC 2616] IETF RFC 2616 (1999), *Hypertext Transfer Protocol – HTTP/1.1*.

[IETF RFC 2818] IETF RFC 2818 (2000), *HTTP Over TLS*.

[IETF RFC 3550] IETF RFC 3550 STD 0064 (2003), *RTP: A Transport Protocol for Real-Time Applications*.

[W3C XML 1.0] W3C XML1.0 (2008), *Extensible Markup Language (XML) 1.0*, (Fifth Edition).

[W3C XML Schema] W3C XML Schema (2004), *XML Schema Part 2: Datatypes Second Edition*, W3C Recommendation 28 October 2004.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 adaptive differential pulse code modulation (ADPCM) [b-ITU-T G.701]: ADPCM algorithms are compression algorithms that achieve bit rate reduction through the use of adaptive prediction and adaptive quantization.

3.1.2 multipurpose Internet mail extensions (MIME) [b-ITU-T J.200]: An application layer protocol. It features a content architecture to facilitate multimedia data such as text other than US-ASCII code, sound, image, etc. to be handled in Internet mails.

3.1.3 pulse code modulation (PCM) [b-ITU-T J.177]: A commonly-employed algorithm to digitize an analog signal (such as a human voice) into a digital bit stream using simple analog-to-digital conversion techniques.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 automatic speech recognition (ASR): A system that can recognize continuous speech, often having phoneme-sized references, using lexical, syntactic, semantic, and pragmatic knowledge, and reacts appropriately (therefore having interpreted the message and found the corresponding action to be taken). [b-ITU-T P.10]

3.2.2 machine translation (MT): Text in a source language is converted by computers into text in a target language which has the same meaning as the original text in the source language.

3.2.3 modality conversion (MC): The conversion of data to different formats/languages using ASR, MT and TTS systems.

3.2.4 modality conversion markup language (MCML): An XML schema that serves as a data description for data exchanged among modality conversion modules.

3.2.5 modality conversion protocol (MCP): The communication protocol which transfers data between MC clients and servers using HTTP(S)/RTP [IETF RFC 2616], [IETF RFC 2818], [IETF RFC 3550]. This protocol transfers the MCML comprised of multimodal information (MI) data which is input into MC clients by users and MC results which are obtained by MC servers.

3.2.6 multimodal information (MI): The information input into MC clients by users via multimodal sensors.

3.2.7 N-best: The most likely "N" hypotheses obtained from modality conversion engines.

3.2.8 speech-to-speech translation (S2ST): Speech in a source language is translated into speech in a target language.

3.2.9 text-to-speech (TTS) synthesis: A process that generates a speech signal from text codes. It is usually composed of the parts: a language-dependent text processing part (the high level processing part), which generates from the character string (by reading rules, vocabulary and semantic analysis) and a set of phonetic, prosodic, etc., parameters that are used by an acoustical signal generating part, the synthesiser itself, which produces the audible speech. [b-ITU-T P.10]

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ADPCM	Adaptative Differential Pulse Code Modulation
ASR	Automatic Speech Recognition
HTTP	HyperText Transfer Protocol
HTTPS	HyperText Transfer Protocol Secure
ID	Identifier
MC	Modality Conversion
MCML	Modality Conversion Markup Language
MCP	Modality Conversion Protocol
MI	Multimodal Information
MIME	Multipurpose Internet Mail Extensions
MT	Machine Translation
PCM	Pulse Code Modulation
RTP	Real Time Protocol
S2ST	Speech-To-Speech Translation
TTS	Text-To-Speech
UCS	Universal Character Set
UTF-8	UCS Transformation Format-8
XML	Extensible Markup Language

5 Conventions

In this Recommendation:

– The expression "is required to" indicates a requirement which must be strictly followed and from which no deviation is permitted if conformance to this Recommendation is to be claimed.

– The expression "is recommended to" indicates a requirement which is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The expression "can optionally" indicates an optional requirement which is permissible, without implying any sense of being recommended.

6 High-level functional model and generic service description

6.1 System overview

The objective of this clause is to describe the high-level functional model of network-based S2ST, in a combination of multiple MC clients and servers. Figure 6-1 shows the functional model for a network-based S2ST.

The functional architecture for network-based S2ST systems requires the following components: S2ST clients and S2ST servers, such as ASR, MT and TTS servers. The S2ST clients are MC clients, which request modality conversion of multimodal information inputted by users to MC servers and which receive the modality conversion results from MC servers. The ASR, MT and TTS servers are MC servers that convert the modality of the multimodal information (MI) input by users.

To achieve the above communication between MC clients and servers, a communication protocol is required, namely the modality conversion protocol (MCP) with the modality conversion markup language (MCML). Based on communication via MCP, MCP clients send MCML to the MCP servers, and MCP servers receive MCML from the MCP clients. The above architecture for the network-based S2ST can also be set up in a standalone system in which MCML is used as an internal interface.

6.2 Functional model of modality conversion (MC) through communication between modality conversion protocol (MCP) clients and servers

The architecture of modality conversion (MC) is shown below in Figure 6-2, which is further elaborated in [ITU-T H.625]. Figure 6-2 shows the functional architecture of MCP clients and servers. This architecture is simply specified in a single MCP client and server combination. Each of the engines for ASR, MT and TTS is identified as a modality conversion function in the MCP server.

The following functions are required within both MCP clients and servers:

– MI processing function;

– MCML processing function;

– communication processing function.

An input and output processing function for MI is required within MCP clients, and a modality conversion (MC) function is required within MCP servers.

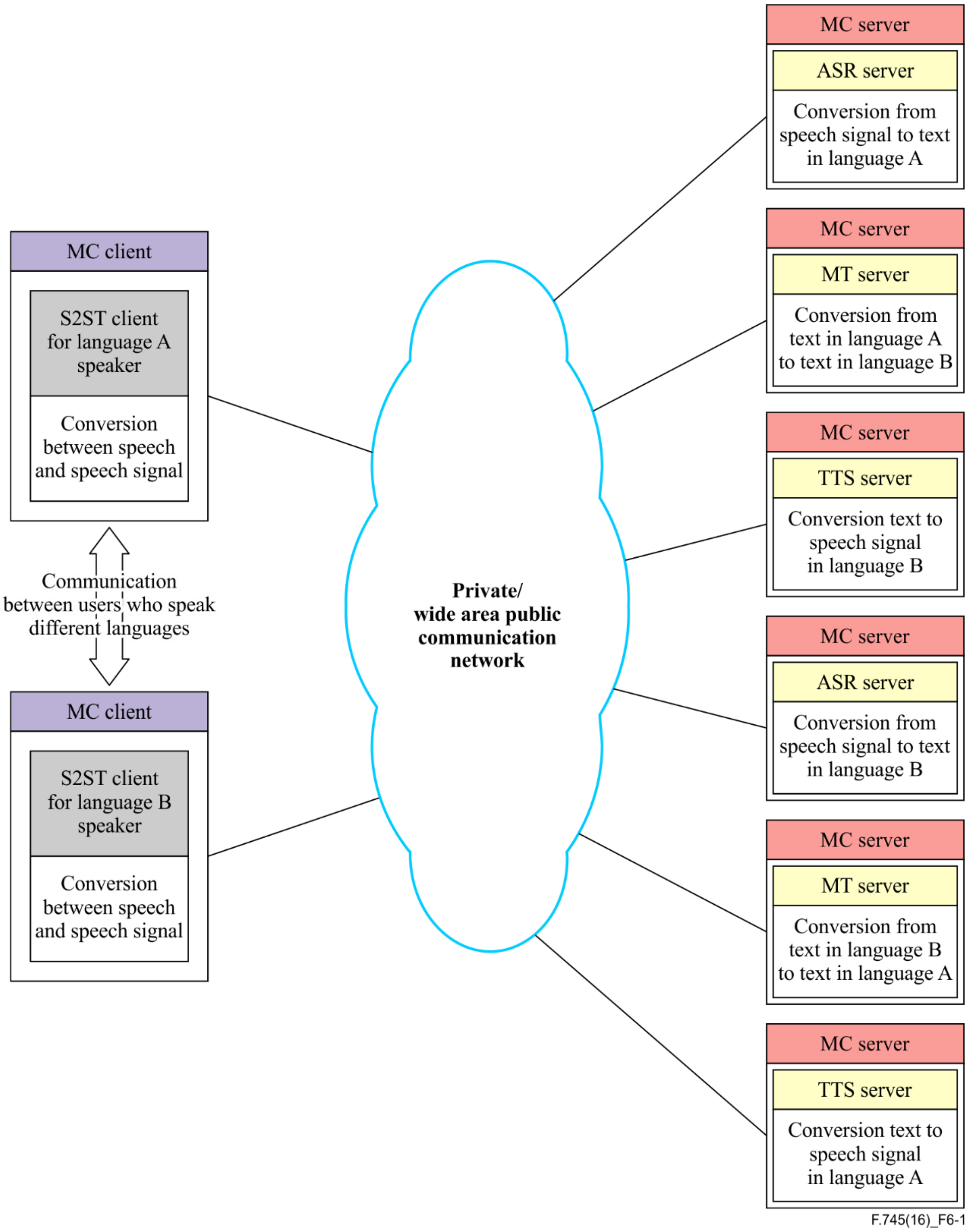
MC server
ASR server
Conversion from speech signal to text in language A
MC server
MT server
Conversion from text in language A to text in language B
MC server
TTS server
Conversion text to speech signal in language B
MC server
ASR server
Conversion from speech signal to text in language B
MC server
MT server
Conversion from text in language B to text in language A
MC server
TTS server
Conversion text to speech signal in language A
MC client
S2ST client for language A speaker
Conversion between speech and speech signal
Communication between users who speak different languages
MC client
S2ST client for language B speaker
Conversion between speech and speech signal
Private/ wide area public communication network
F.745(16)_F6-1

Figure 6-1 – Functional model of a network-based S2ST

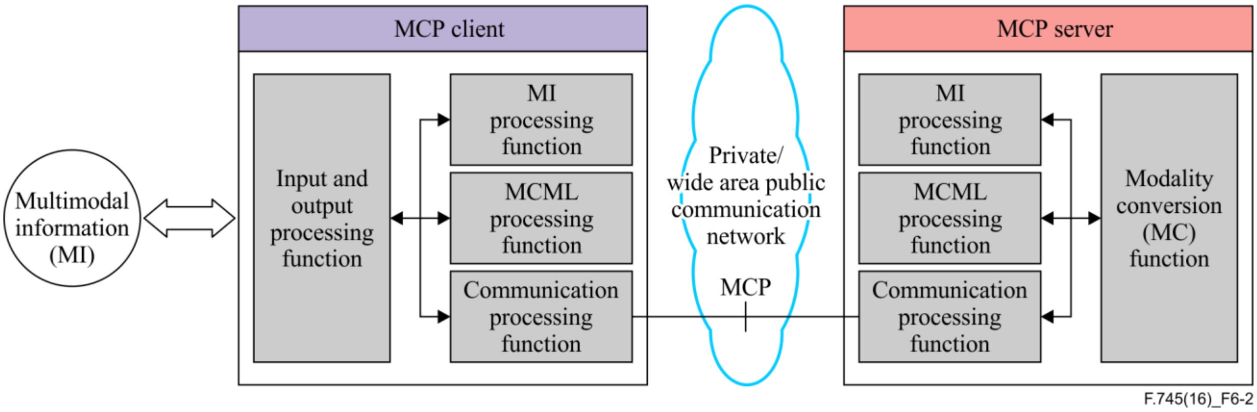
Multimodal information (MI)
MCP client
Input and output processing function
MI processing function
MCML processing function
Communication processing function
Private/ wide area public communication network
MCP
MCP server
MI processing function
MCML processing function
Communication processing function
Modality conversion (MC) function
F.745(16)_F6-2

Figure 6-2 – Architecture of modality conversion clients and servers

6.2.1 Modality conversion markup language

Modality conversion markup language (MCML) is formatted in XML [W3C XML 1.0], which contains the information exchanged between MC clients and servers. UTF-8 [IETF RFC 2279] encoded text is required for MCML.

MCML includes:

- user information of MC clients, such as ID, age and gender;
- device information of MC clients, such as URI [IETF RFC 2396];
- MI input into MC clients by users;
- type of services, device information and order of processes of MC servers;
- MC results obtained from MC servers;
- history of MC services.

6.2.2 Modality conversion protocol

Modality conversion protocol (MCP) is a communication protocol based on HTTP [IETF RFC 2616] or HTTPS [IETF RFC 2818]. To handle real-time streaming, RTP [IETF RFC 3550] is used as an option. MCML and binary data are embedded into the multipurpose Internet mail extensions (MIME) format for transfer.

6.2.3 Multimodal information

MCML is required to be expandable for more complicated modality conversion, to achieve communication using multiple modalities. Information input from users using multiple modalities is defined as multimodal information (MI).

6.2.4 MCP client

Clients that communicate via MCP are denoted as MCP clients and have the following functions:

- send requests formatted in MCML to MCP servers; and
- receive MCML, including MC results, that are sent by MCP servers.

6.2.5 MCP server

Servers that communicate via MCP are denoted as MCP servers and have the following functions:

- receive requests formatted in MCML that are sent by MCP clients; and
- send MCML data, including MC results, to MCP clients.

6.2.6 MC client

A modality conversion (MC) client operates by requesting modality conversion of the multimodal information (MI) which is input by users. The MC client has two different functions:

- transfer function for user and device information, and MI input by users; and
- receive function for MC results from MC servers.

When different devices are used for transmitters and receivers, the functions of MC client can be performed by switching between the MCP client and the MCP server modes.

6.2.7 MC server

A modality conversion (MC) server operates by receiving requests for modality conversion of the multimodal information (MI) that comes from MC client users. The MC server contains MC engines, such as ASR, MT and TTS, which enable modality conversion. In the case where an MC server transfers MC results to another MC server directly through an MC relay communication, the sender MC server uses the MCP client mode.

6.3 Service description

The service defined in this Recommendation enables users to leverage network-based S2ST in a pragmatic manner. Appendix I describes three examples of network-based S2ST services. This service is generically realized by the following operations:

1) Speech is input by users to S2ST clients.
2) S2ST clients send device and speech information to ASR servers.
3) Speech in source language is recognized by ASR servers.
4) The recognized speech is encoded by ASR servers and transferred to MT servers.
5) The encoded speech is translated from the source language into the target language by MT servers.
6) The translation results are transferred to TTS servers.
7) The translation results are converted into speech in the target language by TTS servers.
8) The speech generated by TTS is transferred to S2ST clients.

The MI processing function encodes the MI digitized by the input and output processing function into feature expressions. This function can be implemented either in the MCP client or the MCP server.

7 Requirements

This clause provides a list of basic requirements for services of network-based S2ST. Network-based S2ST has the following requirements in terms of user input, network, user device, MC client, MC server, quality, codec, security and privacy.

7.1 User input requirements

USR-001: This service is required to be usable by people who speak different languages, in face-to-face or remote communication scenarios.

USR-002: This service is recommended to be usable by people who communicate in various modalities, including verbal communication and sign language.

7.2 Network requirements

NW-001: This service requires the network to provide bandwidth and speed sufficient to transfer MI (such as speech and video data) between MCP clients and servers to accomplish real-time communication.

7.3 User device requirements

USR device-001: It is recommended that the user device for this service be usable by people who communicate in various modalities, including verbal communication and sign language.

USR device-002: It is required that the user device for this service include input processors to digitize MI, which is generated through the use of sensing tools such as microphones, keyboards, cameras, etc.

USR device-003: The user device for this service can optionally include an encoding function where the encoding method corresponds to what the server designates.

USR device-004: It is required that the user device for this service include the ability to perform the following MCML processing functions:

– embed digitized/encoded MI into MCML format; and

– extract MC results obtained by the MC server from MCML data.

USR device-005: It is required that the communication processor for the user device in this service include the ability to operate as an MCP client to transfer the MCML to MC servers.

USR device-006: It is required that the communication processor for the user device in this service include the ability to operate as an MCP server to receive the MCML from MC servers.

USR device-007: It is required that the user device for this service include output processors to reproduce MI from the MC server output, such as ASR, MT and TTS results through the use of loudspeaker, display, etc.

7.4 Modality conversion (MC) client requirements

MC client-001: It is required that the MC client be capable of handling multiple user profiles in shared client devices.

MC client-002: It is required that the MC client be capable of setting up the order of MC processes.

MC client-003: It is required that the MC client be capable of connecting/routing messages to multiple clients and servers.

MC client-004: It is recommended that the MC client be capable of getting information on the physical location of the MC device.

7.5 Modality conversion (MC) server requirements

MC server-001: It is required that the communication processor for the MC servers in this service include the ability to operate as an MCP server to receive the MCML from MC clients or servers.

MC server-002: It is required that the MC server for this service have the ability of encoding digitized MI. In case the user device optionally carries out the encoding, the MC server is required to accept the encoded results.

MC server-003: It is required that the MC server for this service use MC functions, i.e., ASR, MT, TTS, etc.

MC server-004: It is required that the user device for this service include the ability to perform the following MCML processing functions:

- embed digitized/encoded MI into MCML format; and
- extract MC results obtained by the MC server from MCML data.

MC server-005: It is required that the communication processor for user devices in this service include the ability to operate as an MCP client to transfer the MCML to MC servers.

7.6 Quality requirements

QLT-001: It is required that this service be equipped with enough capability to perform each of the processes of ASR, MT and TTS. Real time processing for S2ST is recommended to realize spontaneous communication between users who use an MC client or MC clients.

7.7 Security and privacy requirements

The requirements of this function are for further study.

7.8 Codec requirements

The requirements of this function are for further study.

Appendix I

Service description in applications

(This appendix does not form an integral part of this Recommendation.)

I.1 Shared speech-to-speech translation (S2ST) client of two-party communication

S2ST provides two-party communication using a shared application terminal. In this service, speech-to-speech translation can be performed for two-party communication. Multi-party communication can also be provided, although processing of translation is not performed in a simultaneous manner, but in sequential steps. This set-up can be used only for face-to-face communication shown in Figure I.1.

I.2 Personal speech-to-speech translation (S2ST) client communication

S2ST provides users with both two-party and multi-party communication using personal application terminals. This service can also be performed remotely if the application is used in two-party communication, where the speech translation results of one user can be distributed to multiple users in multiple languages at the same time. This set-up can be used for both face-to-face and remote communication shown in Figure I.1.

I.3 Cross-modality communication

S2ST provides different modality communication between users. For instance, a person who uses the spoken word can communicate with a counterpart who is using sign language through the S2ST system. Text-input users can also converse with someone who uses sign language.

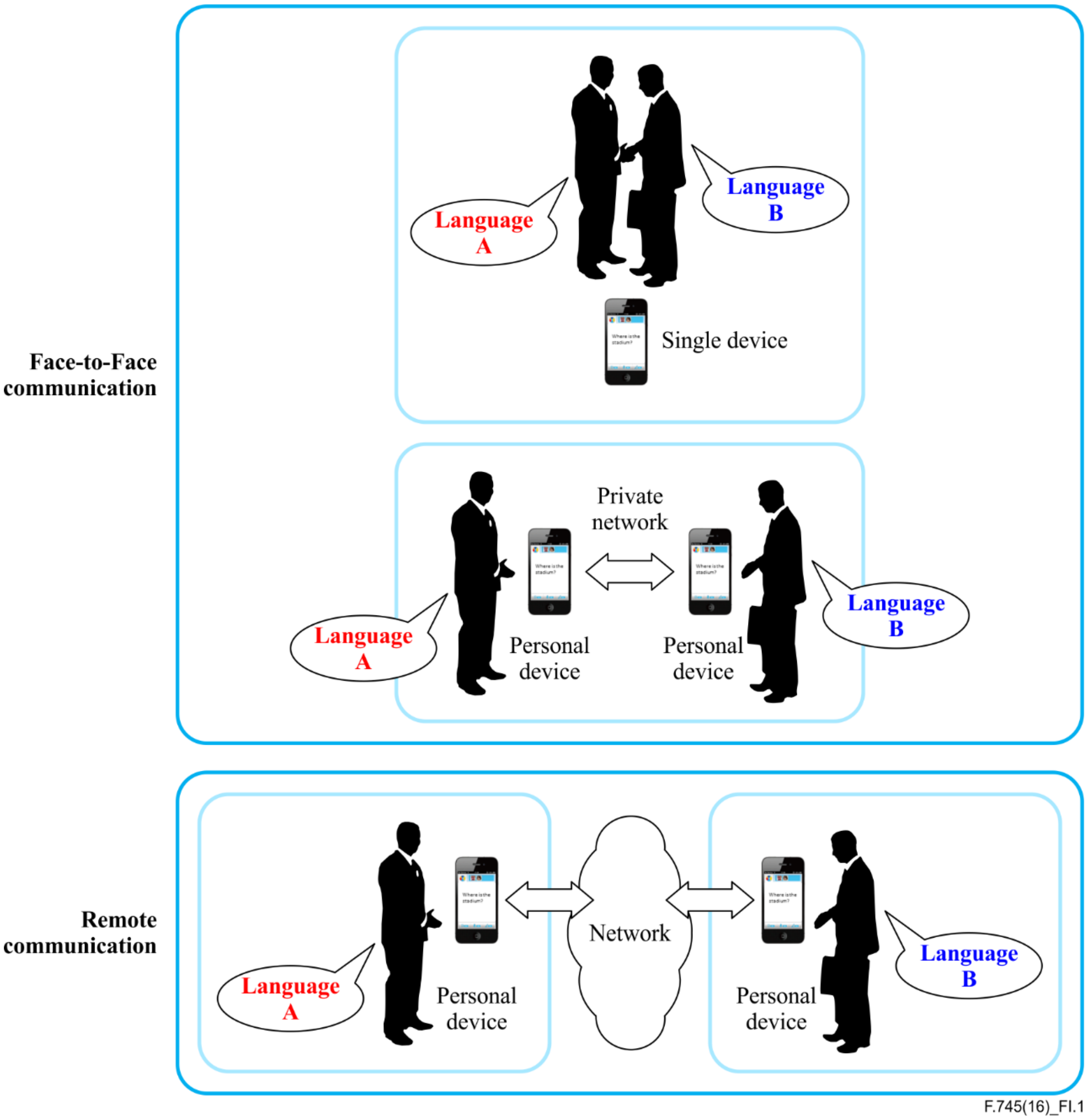

Figure I.1 – Speech translation for face-to-face and remote communication

Bibliography

[b-ITU-T G.701] Recommendation ITU-T G.701 (1993), *Vocabulary of digital transmission and multiplexing, and pulse code modulation (PCM) terms*.

[b-ITU-T J.177] Recommendation ITU-T J.177 (2003), *IPCablecom CMS subscriber provisioning specification*.

[b-ITU-T J.200] Recommendation ITU-T J.200 (2001), *Worldwide common core – Application environment for digital interactive television services*.

[b-ITU-T P.10] Recommendation ITU-T P.10 (1998), *Vocabulary of terms on telephone transmission quality and telephone sets*.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Terminals and subjective and objective assessment methods
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2016