DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 20382-1

ISO/IEC JTC **1**/SC **35** Secretariat: **AFNOR**

Voting begins on:
2016-10-14

Voting terminates on:
2017-01-05

Information technology — User interfaces — Face-to-face speech translation —

Part 1:
User interface

Technologies de l'information — Interface utilisateur — Face-à-face discours traduction —

Partie 1: Interface utilisateur

ICS: 35.240.30

Reference number
ISO/IEC DIS 20382-1:2016(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 20382 was prepared by Joint Technical Committee ISO/IEC JTC 1, Information technology, Subcommittee SC 35, User interfaces.

ISO/IEC 20382 consists of the following parts, under the general title Face-to-Face Speech Translation:

— Part 1: User Interface;

— Part 2: System architecture and functional components;

Introduction

It is important to consider people with special requirements to ensure that they can gain the same benefits from ICT. One of those special requirements is to help people to avoid language barriers in the globalized environments. It has been a long time for automatic speech translation systems existed, but they have functional limitations as well as technical ones with regard to usability and accessibility.

One reason for the limitations is the diversity of the languages currently used. It is difficult to support many languages by one or several speech translation systems. It is required to have a flexible and interoperable standardized framework to work with all different languages utilizing a lot of speech translation systems developed in many countries. Other considerations to make a natural and usable speech translation service possible include applying users' characteristics in the system such as emotion, speech style, gender type and other attributes. To reflect those characteristics in the output speech translation, a standardized user interface is required to reflect the input and output data and transfer them to the user's device.

The main purpose of the International Standard is to help users of different languages to have speech translation service in easier and more convenient ways with the standardized face-to-face speech translation.

Information technology — User interfaces — Face-to-face speech translation —

Part 1:
User interface

1 Scope

This document specifies Face-to-Face Speech Translation designed to interoperate among multiple translation systems with different languages. It also specifies the speech translation features, general requirements and functionality which is a framework to support a convenient speech translation service in the face-to-face situation. The scope includes description of user interface for speech translation and communication protocols for a set-up of the translation session among users. This document excludes defining speech translation engine itself.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 20382-20xx Information technology — User Interface — Face-to-Face Speech Translation – Part 2: System architecture and functional components

3 Terms and definitions

For the purposes of this standard, the following terms and definitions apply.

3.1
Face-to-Face
situation where two users are physically in the same location

3.2
SRWC
(short range wireless communication) wireless communication that uses signals that travel from a few centimeters to several meters

EXAMPLE Bluetooth

4 Abbreviated terms

For the purposes of this standard, the following abbreviated terms apply.

AI

audio indicator

TTS

text to speech

VI

video indicator

WD

wearable device

MD

mobile device

5 Overview of face-to-face speech translation

5.1 General

Face-to-face speech translation system enables users of different languages in a face-to-face situation to communicate each other with spoken languages by providing machine translation results as in Figure 1. In face-to-face speech translation system, mobile devices and wearable devices such as earphones are used for convenient user experiences. The main functions of wearable devices in the translation system are processing input and output of speech signal as a microphone and speakers. Speech recognition and speech synthesis are performed in each user's mobile device. The machine translation function resides in the translation servers.

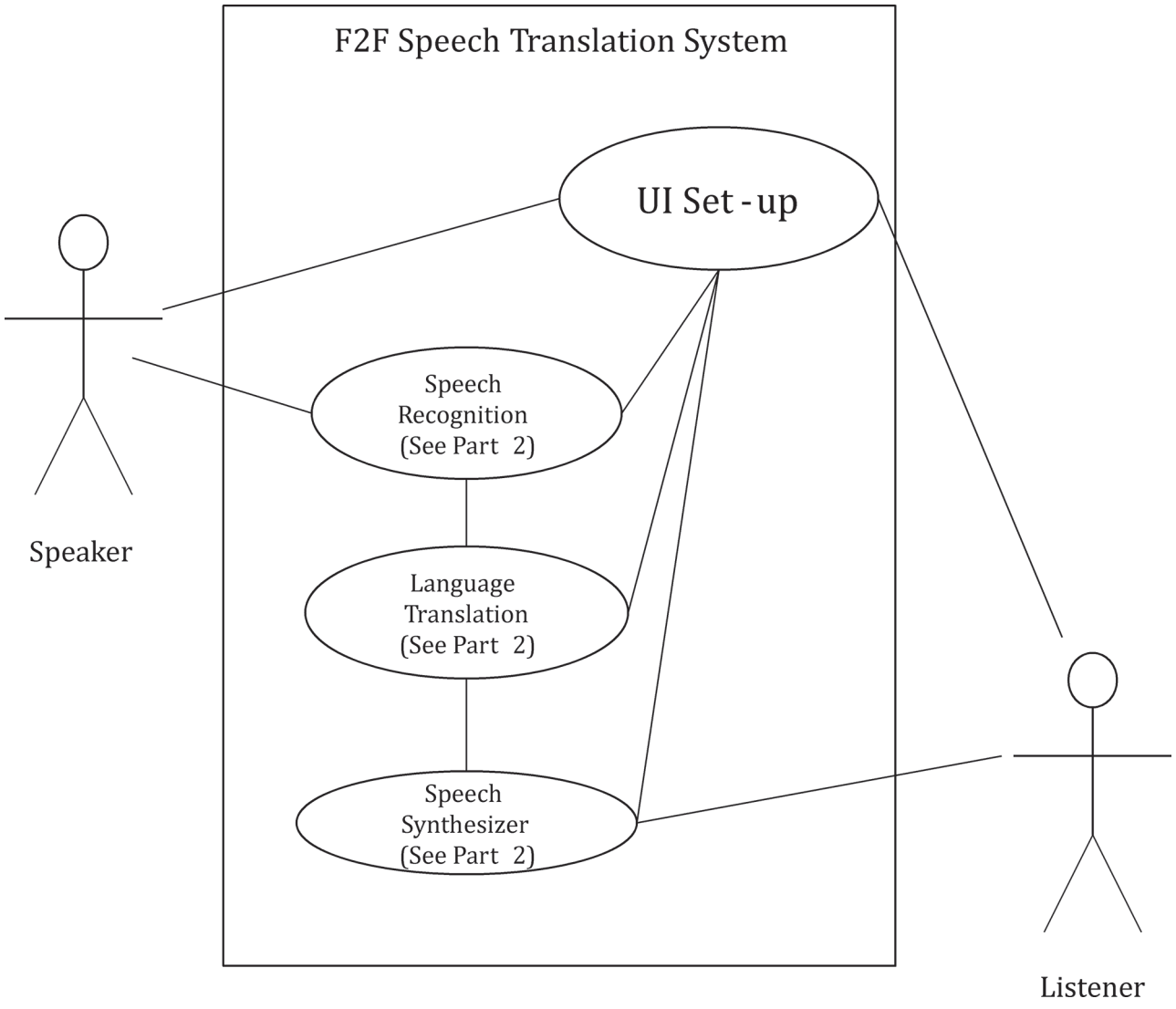


Figure 1 — Service example of face-to-face speech translation

Description of Figure 1 for accessibility purpose:

This figure consists of three vertical parts, i.e., the left part (actor as a speaker), the middle part, and the right part (actor as a listener). In the middle part, there is a box titled 'F2F speech translation system'. The box has three ellipses in a row. They have text "speech recognition (see ISO 20382-20xx part 2)", "language translation (see ISO 20382-20xx part 2)", and speech synthesizer (see part 1), connected with solid lines. On the upper right corner an additional ellipse with a text 'UI set-up' has three solid lines to the three ellipses. The speaker is linked to the UI Set-up ellipse and speech recognition ellipse. On the other hand, the listener is linked to the UI Set-up ellipse and Speech Synthesizer ellipse.

5.2 Service flow

The following steps are general speech translation service processes in face-to-face speech translation system. For more information, see 6.2 and Figure 2 in ISO 20382-20xx Part 2.

1) After a session connection between user A and user B, speech is input by user A in language A. The speech signal is transmitted to the mobile device of user A.

2) The speech is recognized by speech recognition module in the mobile device of user A. Then, the translation operation is requested to the machine translation server K.

3) The translation is performed at server K and the result of it, which is in language B, is sent back to the mobile device of user A in a text form.

4) The translation result is now sent to the mobile device of user B through SRWC communication.

5) The translation result which is in a text form is transformed to speech signal in language B by a speech synthesizer (TTS: text to speech) and sent to the wearable device of user B.

6) For user B's speech, steps (6) to (10) are performed in the same way and the session ends

5.3 Service types

There are two service types depending on the number of participants in the translation session:

— Two way translation: Two users are participating in the translation session.

— Multi-way translation: More than three users are participation in the translation session. The participants may start the dialogue at a different point of time.

5.4 Service mode

Several service modes are selected for usability in different situations.

— Open mode: In this mode, the dialogue is not protected and can be heard to the public. Any user can barge in the dialogue.

— Protected mode: In this mode, only allowed users can participate in the translation session and privacy is guaranteed. The dialog is protected and not heard to the public.

— Automatic mode: The session starts and ends by automatic operation of WD.

— Manual mode: In this mode, the user can decide to be connected to other users manually by his intention.

5.5 Service situation

Several situations are classified as crowded and non-crowded depending on the number of candidates of the user.

— Crowded situation: In this situation, many candidates are around for the translation session.

— Non-crowded situation: In this situation, only one or two candidates are around for the translation session.

6 Functional Requirements

6.1 User Communication Requirements

This sub-clause provides requirements regarding user communication function for face-to-face speech translation:

6.1.1 Required

— The translation system shall allow the users to start a translation session with less than 3 operations

Note The user is able to start a translation session with as limited number of operations as possible.

— The translation system shall allow the users to start a translation session within 10 seconds.

6.1.2 Optional

— The translation system should allow the users to have a session with multiple users.

— The translation system should allow the users to have additional participants after the session starts.

— The translation system should allow the users to have a session with available target users by approaching to them.

6.2 User Interface Requirements

This sub-clause provides the requirements regarding user interface for face-to-face speech translation:

6.2.1 Required

— There shall be no restriction on the users to use the translation system.

Note Any user can use the UI of the translation system.

— The results of speech recognition and translation shall be displayed on the screen of MD.

6.2.2 Optional

— The user should be able to edit the text output of the speech recognition.

— Frequently used functions should be shown on the top-level menu.

— The depth of menu should not exceed 4 steps.

— There should be a button to go back to the top-level menu.

— The user interface should provide options to select alternate translation results.

Note The user should be able to choose the translation results which are speech or text on the screen.

— The user interface should provide options for preferences such as gender, emotion, speech style and other features considered in the futures defined in the user profile.

— The user interface should provide the functionality that reflects the user's characteristics defined in the user profile.

6.3 User Device Requirements

This sub-clause provides requirements regarding user devices used for face-to-face speech translation:

6.3.1 Required

None

6.3.2 Optional

— The translation device should provide a function to allow the user to show the signal to participate in the current translation session.

— The translation devices should to be operated easily.

Note Anyone should be able to operate the devices without difficulties. For example, if the time to start operating a specific function takes more than 3 minutes, it does not satisfy this requirement.

6.4 Accessibility Requirements

This sub-clause provides requirements regarding accessibility function for users with disabilities and specific needs:

6.4.1 Required

— The input from the user shall be possible in a text form as well as a speech form for people with speaking disabilities.

— The translated results shall be provided in a text form as well as speech for people with hearing disabilities.

6.4.2 Optional

— The translated results should be provided also in a controlled language for people with mental disabilities.

Note Controlled languages are defined in ISO TC37 (ISO/TS 24620-1:2015 Language resource management — Controlled natural language (CNL) — Part 1: Basic concepts and principles)

7 Functional Components of face-to-face speech translation

7.1 Service protocol among functional components

The workflow and service protocol among components of face-to-face speech translation are described in Figure 2:

1. Connection stage:

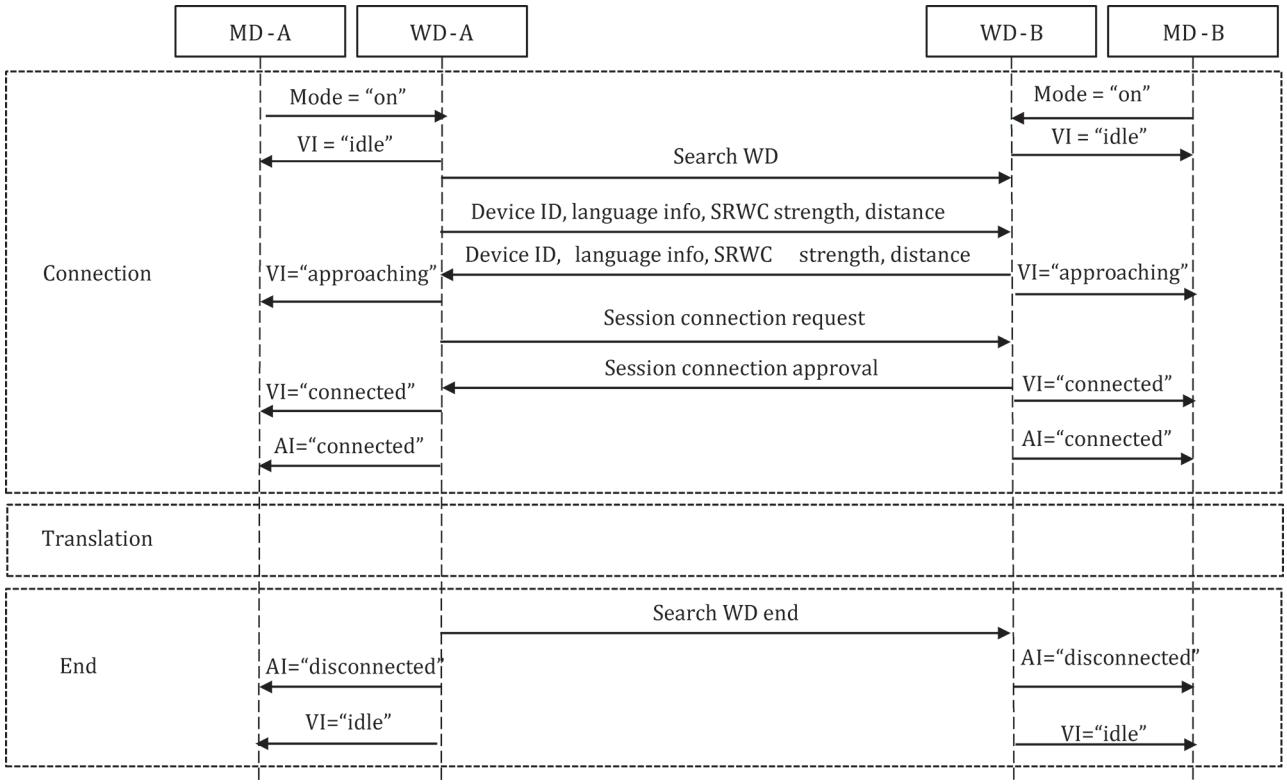
   In the connection stage, the users who wish to participate in the translation service set the translation mode as “ON” state and this information is sent to each user’s wearable device (WD). Then the WD sets the Visual Indicator (VI) as “idle” meaning the user is in the translation ready mode. WD automatically starts searching for available WDs nearby and exchanges device ID and language information with them. It then measures the signal strength of the WDs and produces a candidate WD target list. The WD searches for the direction/approaching information of the candidate WDs, and sends the session connection message with the Visual Indicator “approaching”. When the WD connection is approved, Visual Indicator of both WDs turn into “connected” state with Audio Indicator (AI) signalling “connected”.

2. Translation stage:

   After the connection between two WDs is made, the translation starts. The translation stage will be described in detail in ISO 20382-20xx Part 2 of the standard.

3. End (disconnection) stage:

   The WD checks regularly if the session needs to be disconnected. If it finds no activity it turns the Visual Indicator to “idle” with Audio Indicator “disconnected” and ends the session. It also resets the target languages.



- MD - A: Mobile device of language A
- WD - A: Wearable device of language A
- MD - B: Mobile device of language B
- WD - B: Wearable device of language B

Figure 2 — The sequence diagram of face-to-face speech translation (UI set-up)

Description of Figure 2 for accessibility purpose:

The figure consists of four horizontal layers. The top layer shows 4 nodes for information flows. On top, there are 4 boxes for 4 nodes in a row with a vertical dotted line connected to each box. The names are ‘MD-A’, ‘WD-A’, ‘WD-B’ and ‘, MD-B’. The names stand for ‘mobile device of language A’, ‘wearable device of language A’, ‘wearable device of language B’ ‘mobile device of language B’, respectively.

Under the top layer there are three layers and each layer is in a box: Connection layer, Translation layer and End layer.

The three layers under the top layer represent the information flow among the nodes on the top layer.

First, in the connection layer, there are arrows among the nodes in the following order:

1. A rightwards arrow from MD-A goes to WD-A with a label “Mode setting = ON”

At the same time: a leftwards arrow from MD-B goes to WD-B with a label “Mode setting = ON” (This is the first line in the connection layer)

2. A leftwards arrow from WD-A goes to MD-A) with a label “VI = “idle””

At the same time: a rightwards arrow from WD-B goes to MD-B with a label “VI = “idle”” (This is the second line in the connection layer)

3. A rightwards arrow from WD-A goes to WD-B with a label “Search WD”

4. A rightwards arrow from WD-A goes to WD-B with a label “Device ID, language info, SRWC strength, distance”

5. A leftwards arrow from WD-B goes to MD-B with a label “Device ID, language info, SRWC strength, distance”

6. A leftwards arrow from WD-A goes to MD-A with a label “VI = “approaching””

At the same time: a rightwards arrow from WD-B goes to MD-B) with a label “VI = “approaching””

7. A rightwards arrow from WD-A goes to WD-B with a label “Session connection request”

8. A leftwards arrow from WD-B goes to WD-A with a label “Session connection approved”

9. A leftwards arrow from WD-A goes to MD-A with a label “VI = “connected””

At the same time: a rightwards arrow from WD-B goes to MD-B with a label “VI = “connected””

10. A leftwards arrow from WD-A goes to MD-A with a label “AI = “connected””

At the same time: a rightwards arrow from WD-B goes to MD-B with a label “AI = “connected””

Next, in the translation layer, there is no information since it is described in ISO 20382-20xx Part 2.

The third layer is the End layer shows the flows in the following order.

1. A rightwards arrow from WD-A goes to WD-B with a label “Search WD end”

2. A leftwards arrow from WD-A goes to MD-A with a label “AI = “disconnected”

At the same time: a rightwards arrow from WD-B goes to MD-B with a label “AI = “disconnected”

3. A leftwards arrow from WD-A goes to MD-A with a label “VI = “idle””

At the same time: a rightwards arrow from WD-B goes to MD-B with a label “VI = “disconnected”

7.2 User communication functional block

7.2.1 General

The user communication functional block includes the Connection stage and End stage described in Figure 2. Before the actual translation begins, the user should find a translation partner among candidate users according to the following communication steps.

7.2.2 Steps for user communication

To start the user communication, the user of the translation application first sets up the mode as he wishes. For example, the user may want to use a protected mode, in which only the participants in the translation session can hear the conversation. The next step, as in Figure 3, is the start of the translation session. The translation step itself is between the session start and session end. More details of the translation session connection and disconnection are described in section 2 and 3.

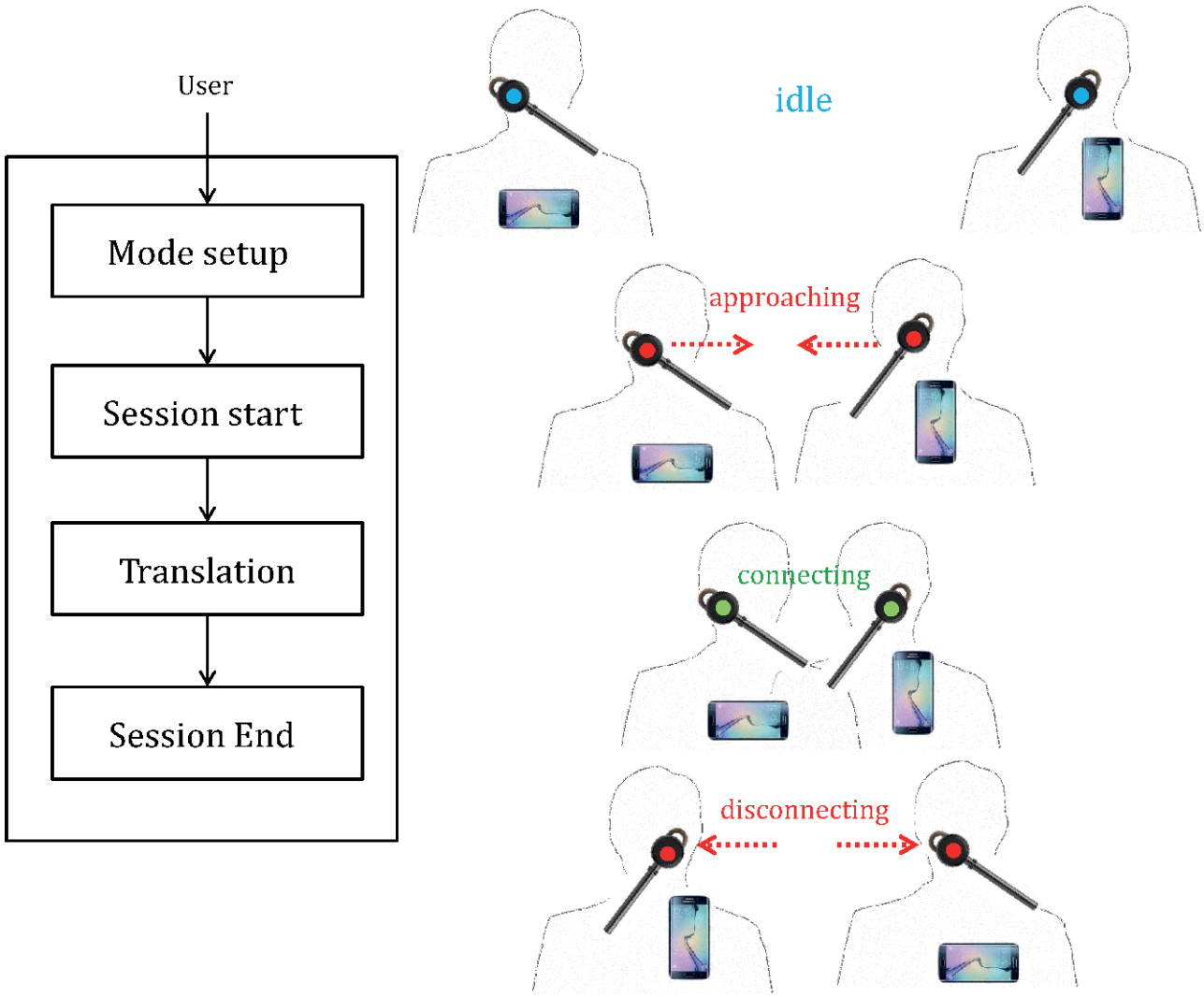


Figure 3 — User communication for translation session

Description of Figure 3 for accessibility purpose:

On the left side of the figure, there is a flow chart which consists of a big box which contains four small boxes connected with downwards arrows. On the top which is outside of the box, there is a “User” and downwards arrow from “user” to the first small box with a label “Mode setup”. Each small box are linked with downwards arrows that signals the flow chart. The second small box is labelled as “Session start”. The third small box is labelled as “Translation”. The fourth one is labelled as “Session End” and there is no more downwards arrow from this box.

On the right side of the figure, there are four horizontal layers with two people in each layer. They have a mobile device and wearable device with them. The first layer is an “idle” layer where two people are in a distance. The second layer is an “approaching” layer, where two people are closer than the first layer. There are also arrows pointing each other signalling they are approaching each other. The third layer is a “connecting” layer where two people are very close. Finally in the fourth layer which is a “disconnecting” layer, the space between two people is wide and the arrows are pointing their face meaning they are departing.

7.2.3 Translation session connection

The translation session starts by searching WD around the user. As shown in Figure 4, the user starts searching WDs by exchanging device ID information, language information, SRWC strength calculation. Then the user produces candidate WD list based on the information he gathered from the WDs around

him. After measuring WD approaching information, WD connection with a best candidate is approved. The decision on the approaching WD is made based on the distance and direction of the candidate. After the translation session connection is made, the translation starts between the two users.

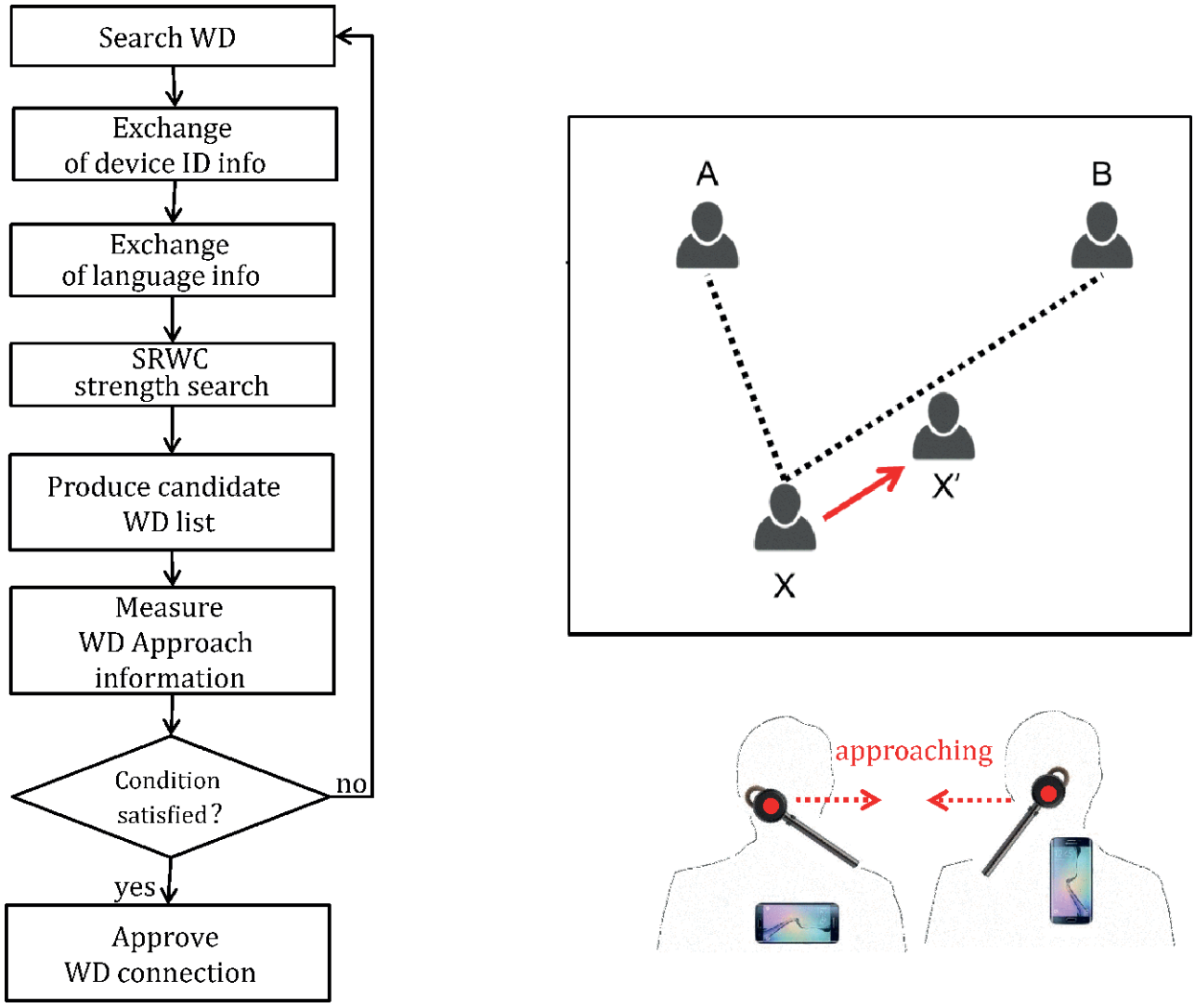


Figure 4 — Translation session connection

Description of Figure 4 for accessibility purpose:

On the left part of the figure, there is a flow chart which consists of 6 rectangular boxes and then a diamond shape and an additional rectangular box connected with downwards arrows. The texts in the boxes and the diamond are written in the following order: 1. Search WD, 2. Exchange of device ID info, 3. Exchange of language info, 4. SRWC strength search, 5. Produce candidate WD list, 6. Measure WD Approach information, 7. Condition satisfied? (in a diamond shape) 8. Approve WD connection (the last box). On the downwards arrow between the diamond and the last box, there is a text “yes”. There is upwards arrow starting from the diamond pointing to the first box (Search WD) with a text “no”.

On the upper right part of the figure, there is a big box and 4 people in it. Two people on the top are marked with “A” and “B” respectively. There is one person in the middle marked with “X’”. The person on the bottom is marked with “X”. The person A, and X are connected with dotted line. The person B and X are also connected with dotted line. There is an arrow from X pointing to X’.

On the bottom right part of the figure, there are two people with a text “approaching” and arrows pointing each other signalling they are approaching to each other. They have a mobile device and wearable device with them.

7.2.4 Translation session disconnection

The disconnection of the translation session is also performed automatically from the user's point of view. When there is no connected WD and session end condition is met, the target language is reset for the next translation session and the session ends.

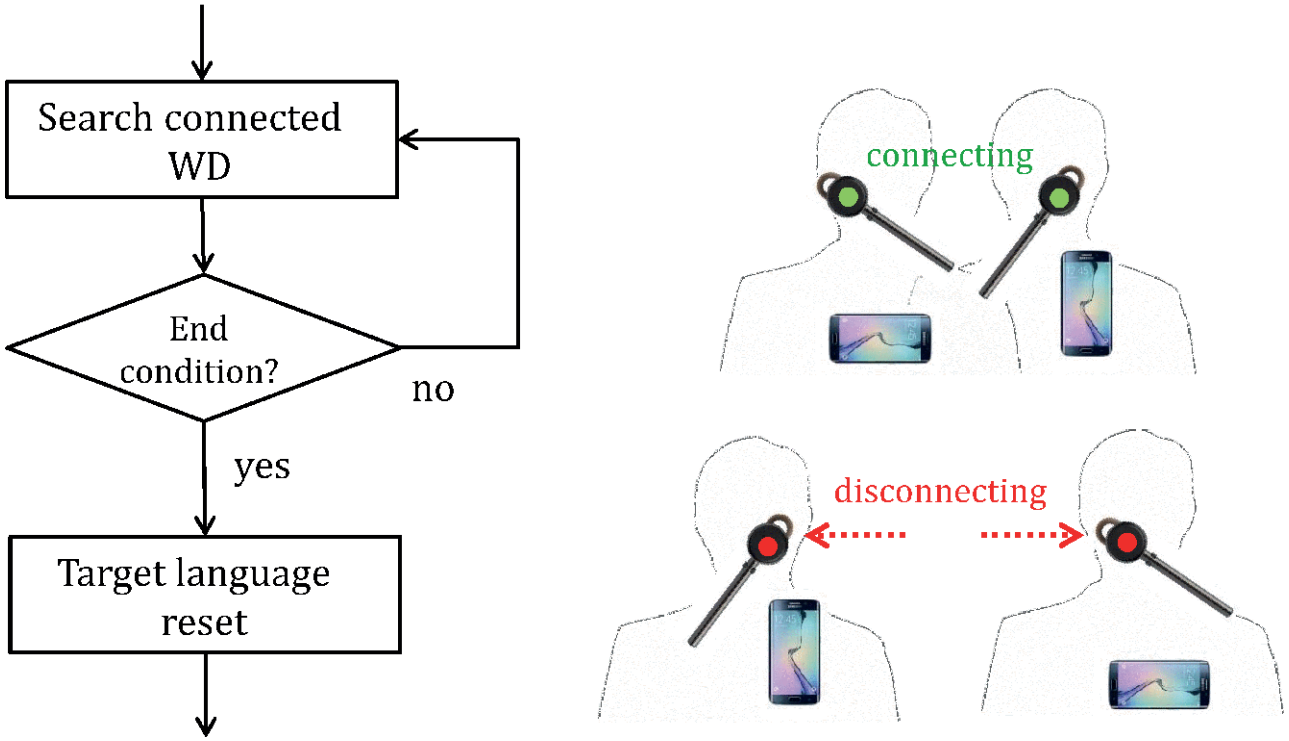


Figure 5 — Translation session disconnection

Description of Figure 5 for accessibility purpose:

On the left part of the figure, there is a flow chart which consists of a rectangular box and then a diamond shape and an additional rectangular box connected with downwards arrows. There is additional downwards arrow from outside into the first box and there is another one from the last box down to outside. The texts in the boxes and the diamond are written in the following order: 1. Search connected WD (the first box), 2. End condition (diamond) 3. Target language reset (the second box). On the downwards arrow between the diamond and the last box, there is a text "yes". There is upwards arrow starting from the diamond pointing to the first box (Search connected WD) with a text "no".

On the upper right part of the figure, there are two people side by side with a text "connecting". They have a mobile device and wearable device with them.

On the bottom right part of the figure, there are two people with a space between them with a text "disconnecting" and arrows pointing themselves signalling they are going away from each other. They have a mobile device and wearable device with them.

7.2.5 Translation session end condition

The translation session end condition is presented in Figure 6, where the existence of translation results, departing WDs and manual end request are one of the ending conditions. When the translation results are not produced for a certain time, the system checks if the WDs are departing or not. In case the translation is continuing but there is a manual end request made, the end condition is still met and the session end is approved.

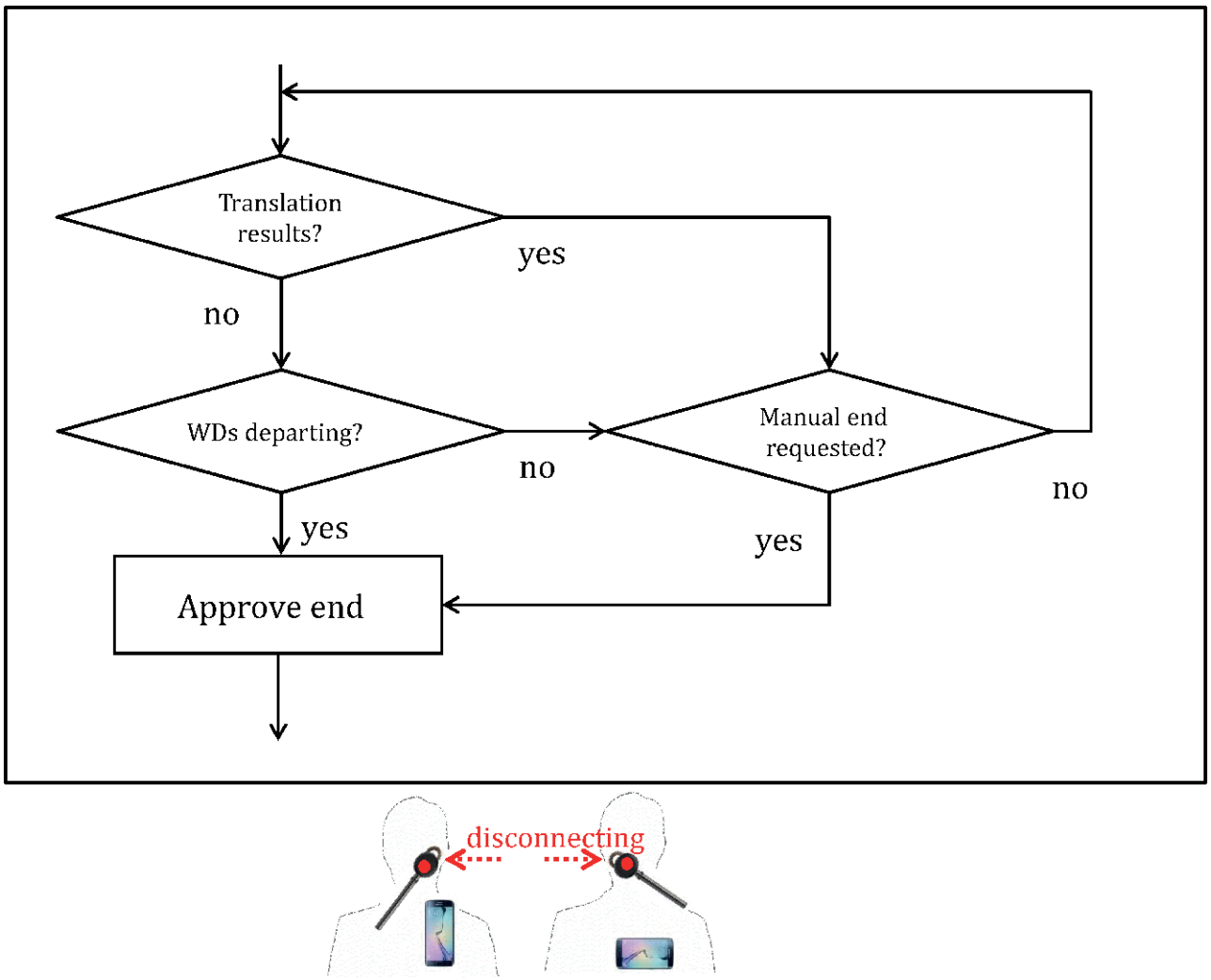


Figure 6 — End condition of translation session

Description of Figure 6 for accessibility purpose:

On the upper part of the figure, there is a flow chart which consists of 3 diamond shapes and a rectangular box connected with downwards arrows. There is an additional downwards arrow from outside into the first diamond and there is another one from the last box down to outside. The first diamond is connected with the second diamond with a downwards arrow marked with “no”. The text of the first diamond is “Translation result?” And that of the second diamond is “WDs departing?” The second diamond is connected to the last box with downwards arrow marked with “yes”. The text of the last box is “Approve end”. The third diamond with the text “Manual end request?” is connected to other diamonds and the last box. It is connected to the last box with downwards arrow marked with yes. The first diamond is connected with the third diamond with a downwards arrow marked with “yes”. The second diamond is connected with the third diamond with a downwards arrow marked with “no”. There is another arrow coming from the third diamond marked with “no” which connects back to the very top arrow from outside.

On the bottom part of the figure, there are two people with a small space between them with a text “disconnecting” and arrows pointing themselves signalling they are going away from each other. They have a mobile device and a wearable device with them.

7.3 User Interface functional block

User interface functional block has the following functions

7.3.1 Setup of the initial translation environments

— Turn on/off of SRWC

— Start of the F2F translation session

— Selection of the source language

— Mode selection between protected/non-protected

7.3.2 Correction/selection function of speech recognition results:

When the user finds errors in the recognized speech, he can correct the sentence by the following methods. In addition the recognized sentence can be provided as several candidate results among which the user selects the correct sentence.

— Keyboard correction

— Gesture correction

— Re-speaking of the error sentence

— Selection of the correct sentence among candidates

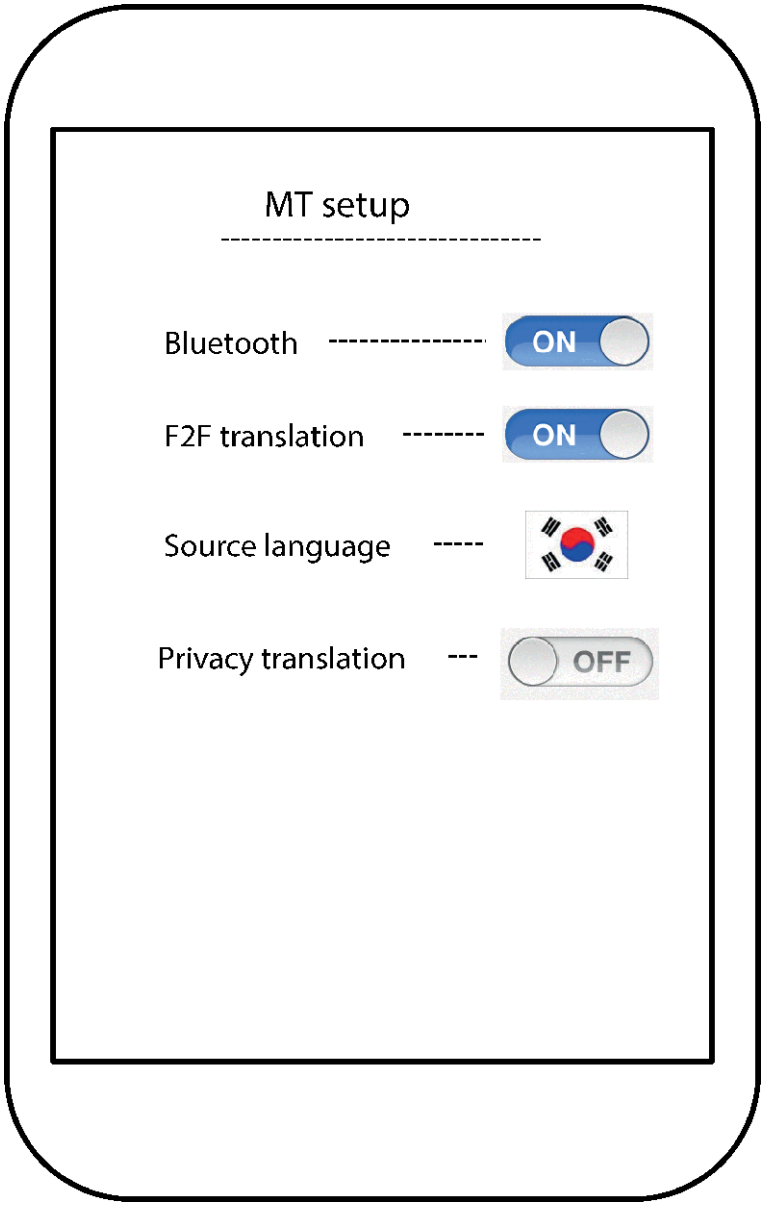


Figure 7 — Setup of user interface functions

Description of Figure 7 for accessibility purpose:

The figure shows the setup interface for the user. There is a rectangular with round corners and a regular rectangular inside it signalling a mobile device. There are five lines which indicates menu options that a user can select. The top line is the title “MT setup”. The second line is “Bluetooth” with a selection button for “on/off”. The third line is “F2F translation with selection button for “on/off”. The fourth line is “Source language” with Korean flag selected. The last line is “Privacy translation” with a selection button for “on/off”. Currently the top two selection button are selected for “ON” and the last button for “OFF”.

7.4 User Device functional block

User device functional block includes the following functions:

7.4.1 SRWC pairing function for search for users and Visual Indicator control

7.4.2 Wearable device functions

7.4.3 Communication between devices and message representation

7.5 Accessibility functional block

Accessibility functional block provides the following functions for people with disabilities and specific needs.

7.5.1 Multi-modal input functions

The multi-modal input functions include text/touch/gesture input for hearing impaired people. People with hearing disabilities will use, for example, text to setup the translation application according to their preferences. They can also input their speech source language using text keyboard that will be translated through the application.

7.5.2 Multi-modal output functions

The multi-modal output functions include text/touch/gesture input for hearing impaired people. People with hearing disabilities will receive the translation results as text instead of speech so that they can read and understand what the other party said in foreign language.

7.5.3 Voice input and output functions

The voice input and output functions are also parts of multi-modal user interface. Other than the intrinsic speech input and output of the translation system, voice-enabled session setup function is provided for blind people and body impaired people.

7.5.4 Accessibility functions for the application UI

The accessibility functions for the application UI include color change, font enlargements for the people with visual difficulties and the elder people.

7.5.5 Simple language, or controlled language function

The simple language, or controlled language function (defined in ISO/TS 24620-1:2015) provides simple output sentences which are easy to understand for people with mental disabilities.

Annex A
(informative)

Comparison of standardization activities for speech translation

In the following table, the differences and similarities of each standardization activity are presented. Three standardization activities are described: SC35, ITU-T SG2, ITU-T SG16

	ISO JTC1 SC35 Culture and Linguistic Adaptability(WG5)	ITU-T SG2 Human Factors (Q4)	ITU-T SG16 Multimedia Service (Q21)
Title	Face-to-Face Speech Translation	User Interface for face-to-face speech translation considering human factors	Network-based speech-to-speech translation services
Scope	Framework for face to face speech translation including devices, servers, interface and communication protocols related to User Interface	User interface features for usability	Functional requirements, service description, functional architecture, mechanism, interface protocol, workflow
Focus	Human to Computer Features	Human factors, Ergonomics, Accessibility	System factors
State	Proposed for CD	Draft Recommendation	Standards published (2010, F.745, H.625)
			Amendment started for face-to-face situation
Project leader	ETRI	ETRI	NICT

Bibliography

[1] ISO/IEC/TR 11017:1998, *Information technology — Framework for internationalization*.

[2] ISO/IEC 30122, *Information technology — User interfaces — Voice commands*.

[3] ISO TC37 (ISO/TS 24620-1:2015 Language resource management — Controlled natural language (CNL) — Part 1: Basic concepts and principles