DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 20382-2

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

Voting begins on:
2016-10-14

Voting terminates on:
2017-01-05

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

Part 2:
System architecture and functional components

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

Partie 2: Architecture du système et des composants fonctionnels

ICS: 35.240.30

Reference number
ISO/IEC DIS 20382-2:2016(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 ..... **iv**

Introduction ..... **v**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**

4 Overview of F2F speech translation **1**
4.1 Genaral 1
4.2 Functional components of F2F speech translation 2

5 Functional Requirements **3**
5.1 General requirement 3
5.2 Speech recognition requirements 3
5.3 Language translation requirements 3
5.4 Speech synthesizer requirements 4

6 System architectures of F2F speech translation **4**
6.1 General 4
6.2 Two persons with embedded F2F speech translation devices 5
6.3 Two persons with remote speech translation functions 7
6.4 Mixture of 6.1 and 6.2 9
6.5 Adding one more speaker to F2F speech translation conversation 11
6.6 Two person with only one fixed F2F speech translation device 13

Annex A (informative) **History of F2F speech translation** **16**

Annex B (informative) **An example scenario of F2F speech translation protocol** **22**

Bibliography **23**

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.

This standard aims to enable face-to-face speech translation among peoples with different languages. The three technologies, i.e., speech recognition, language translation, and speech synthesis technologies, are mature enough to build speech translation function. There are many face-to-face speech translation devices and/or services using mobile devices. However, one should learn how to use the service, and should use both hands to control the speech translation system. If one is using his one hand which is very usual case, it is impossible to use current speech translation systems and/or services. To overcome this usability issue, this Internal Standard suggests a method that exactly follows the conversation among the peoples with the same language. The method in this International Standard is hands-free, and do not need any pre-training. In this sense, this method is an ultimate user interface of face-to-face speech translation, and will open a world without language barriers.

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

Part 2:
System architecture and functional components

1 Scope

This document specifies functional components of 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 speech translation devices, servers, and communication protocols among speech translation servers and clients in a high-level approach. This International Standard also defines various system architectures in different environments. This document excludes defining speech recognition engines, language translation engines, and speech synthesis engines.

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, *Information technology — User Interface — Face-to-Face Speech Translation – Part 1: User Interface*

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

3 Terms and definitions

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

3.1
Utf-8
standard defined in IETF RFC 2279 (1998), UTF-8, a transformation format of ISO 10646

3.2
TTS
SOURCE: ISO/IEC 20382-1xx, 3.X.

4 Overview of F2F speech translation

4.1 Genaral

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. Face-to-face speech translation system between a speaker and a listener shall have a speech recognition module, language translation module, and a speech synthesizer (TTS: text to speech) as shown in Figure 1.

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 are speech recognition, language translation, and speech synthesizer, connected with solid lines. On the upper right corner an additional ellipse with a text 'UI set-up (see ISO 20382-20XX part 1)' 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.

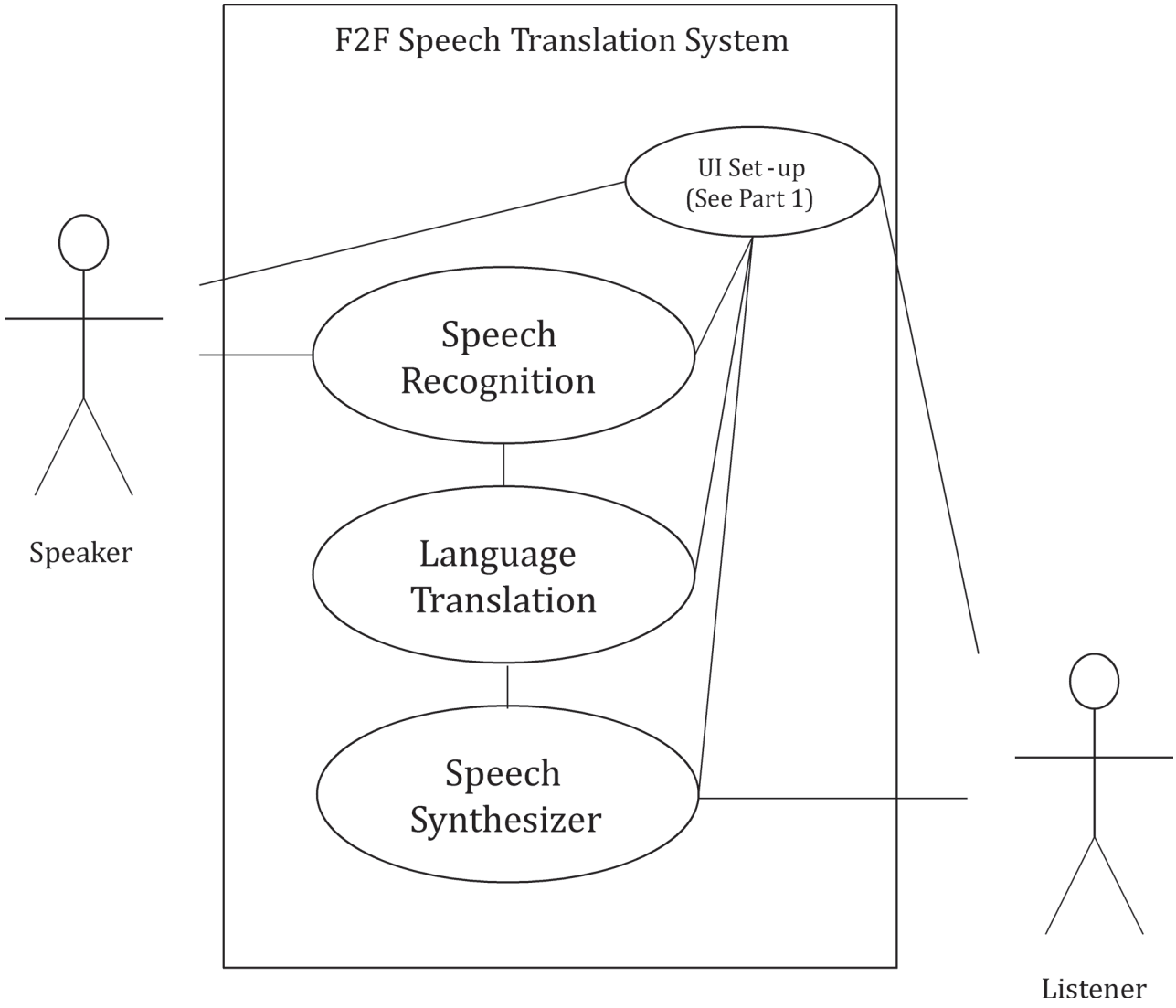


Figure 1 — Functional components of F2F speech translation

4.2 Functional components of F2F speech translation

For F2F speech translation, the speaker and the listener shall set up UI (see ISO 20382-20XX part 1).

The functions of each component in Fig. 1 are as follows.

1) The speaker speaks a sentence in his/her own language.

2) The Speech recognition module recognizes the speech, and outputs the corresponding text.

3) The text is translated into another language of the same meaning through the language translation module.

4) The speech synthesizer generates the corresponding speech in listener's language based on the translated text.

5) Listening to the speech, the listener answers in his/her own language.

6) The steps (2) to (5) continues until the users accomplish their goals.

5 Functional Requirements

5.1 General requirement

This sub-clause provides general requirements regarding face-to-face speech translation:

— There are three remote services in this standard, i.e., remote translation service, remote speech recognition service, and remote speech synthesis service. All these remote services shall keep the privacy of the face-to-face speech translation users.

— The translation system should allow the users to start a translation session as natural as in the everyday conversation.

— The translation system should allow the users to start a translation session as fast as in the everyday conversation (i.e., not exceeding 2 seconds).

— The speech translation system should work in real time (i.e., not exceeding 2 seconds).

— 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 have started.

5.2 Speech recognition requirements

This sub-clause provides the requirements regarding speech recognition module of face-to-face speech translation:

— The speech recognition module shall recognize the speech into the text of the same language.

— The speech recognition module shall accept most popular speech formats.

— The speech format should be defined as a metadata format such as MIME format.

— The output of the speech recognition module should be written in utf-8 format.

— Note: This standard does not specify the data format of the speech nor that of the text since there are many off-the-shelf speech recognition module with various input and output data formats.

5.3 Language translation requirements

This sub-clause provides requirements regarding user language translation module of face-to-face speech translation:

— The language translation module shall translate a text in a source language into a text in a target language with the same meaning.

— If there is no direct language translation module between the source language and the target language, one should use intermediate language to accomplish the language translation. One should translate source language to intermediate language, and then intermediate language to target language. One should choose the intermediate language so that the language translation performance is the best. If there is no performance data available, the intermediate language should be chosen from the

same language family or the languages of the same word order of the source language or the target language.

— Note: This standard does not specify the data formats of the input and output texts since there are many off-the-shelf language translation module with various input and output data formats.

5.4 Speech synthesizer requirements

This sub-clause provides requirements regarding speech synthesizer of face-to-face speech translation:

— The speech synthesizer shall generate the corresponding speech from the text of the same language.

— In face-to-face speech translation the synthesized speech should be as close as possible to the original speaker to increase the naturalness of the conversation. The gender of the synthesized speech in language B should be equal to that of the user in language A. The naturalness can be increased if the base frequency, speed, prosody, and/or speech color of the synthesized speech is similar to those of the original speaker.

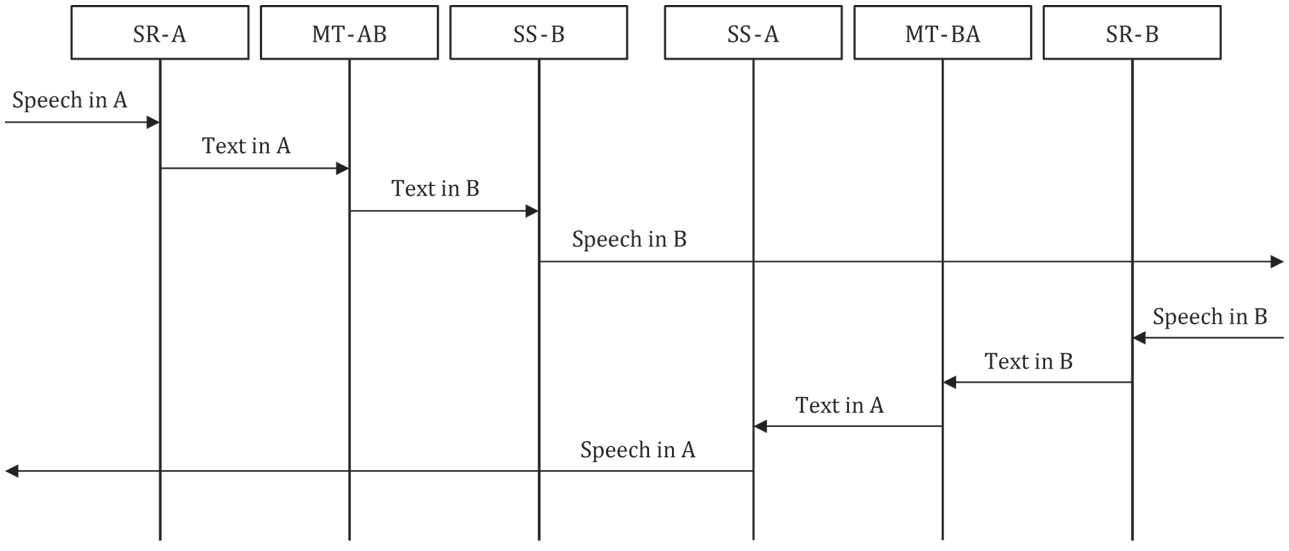
— The text input of the speech synthesizer should be written in utf-8 format.

— Note: This standard does not specify the data format of the speech nor that of the text since there are many off-the-shelf speech synthesizer with various input and output data formats.

6 System architectures of F2F speech translation

6.1 General

Figure 2 shows the sequence diagram of face-to-face speech translation – part 2.



- SR-A: speech recognition of language A
- MT-AB: machine translation from A to B
- SS-B: speech synthesizer of language B
- SR-B: speech recognition of language B
- MT-BA: machine translation from B to A
- SS-A: speech synthesizer of language A

Figure 2 — The sequence diagram

Description of Figure 2 for accessibility purpose:

There are 6 objects in a row with lifelines. The names are ‘SR-A’, ‘MT-AB’, ‘SS-B’, ‘SS-A’, ‘MT-BA’, and ‘SR-B’. The names stand for ‘speech recognition of language A’, ‘machine translation from A to B’, ‘speech synthesizer of language B’, ‘speech synthesizer of language A’, ‘machine translation from B to A’, and ‘speech recognition of language B’, respectively.

There is an arrow with text ‘speech in A’ from the left end to SR-A. Beneath this arrow, there is an arrow with text ‘text in A’ from SR-A to MT-AB. Beneath this arrow, there is an arrow with text ‘text in B’ from MT-AB to SS-B. Beneath this arrow, there is a long arrow with text ‘speech in B’ from SS-B to the right end.

Under the previous arrow, there is an arrow of reverse direction with text ‘speech in B’ from the right end to SR-B. Beneath this arrow, there is an arrow with text ‘text in B’ from SR-B to MT-BA. Beneath this arrow, there is an arrow with text ‘text in A’ from MT-BA to SS-A. Beneath this arrow, there is a long arrow with text ‘speech in A’ from SS-A to the left end.

6.2 Two persons with embedded F2F speech translation devices

The basic system architecture between two persons with embedded F2F speech translation devices is described in Figure 3.

Description of Figure 3 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (wearable device layer), the middle layer (mobile device layer), and the lower layer (translation server layer). This figure can also be divided into two vertical parts, i.e., the left part and the right part. The left part corresponds to the user in language A, and the right part corresponds to the user in language B.

The upper left part shows a wearable device of the user in language A. The middle left part shows a mobile device of the user in language A. There is an arrow from the wearable device to the mobile device with a text of “1) Speech input”. The lower left part shows a cloud shape with a text of “Translation server K” that indicates language translation server of the user in language A. There is an arrow from the mobile device to the translation server K with a text of “2) Translation request”. There is another arrow from the translation server K to the mobile device with a text of “3) Translation result”.

The middle right part shows a mobile device of the user in language B. There is an arrow from the mobile device of the user in language A to the mobile device of the user in language B with a text of “4) Translation result”. The upper right part shows a wearable device of the user in language B. There is an arrow from the mobile device of the user in language B to the wearable device of the user in language B with a text of “5) TTS in B”. There is another arrow from the wearable device of the user in language B to the mobile device of the user in language B with a text of “6) Speech input”. The lower right part shows a cloud shape with a text of “Translation server G” that indicates language translation server of the user in language B. There is an arrow from the mobile device of the user in language B to the translation server G with a text of “7) Translation request”. There is another arrow from the translation server G to the mobile device of the user in language B with a text of “8) Translation result”. There is an arrow from the mobile device of the user in language B to the mobile device of the user in language A with a text of “9) Translation result” and “SWRC”. There is an arrow from the mobile device of the user in language A to the wearable device of the user in language A with a text of “10) TTS in A”.

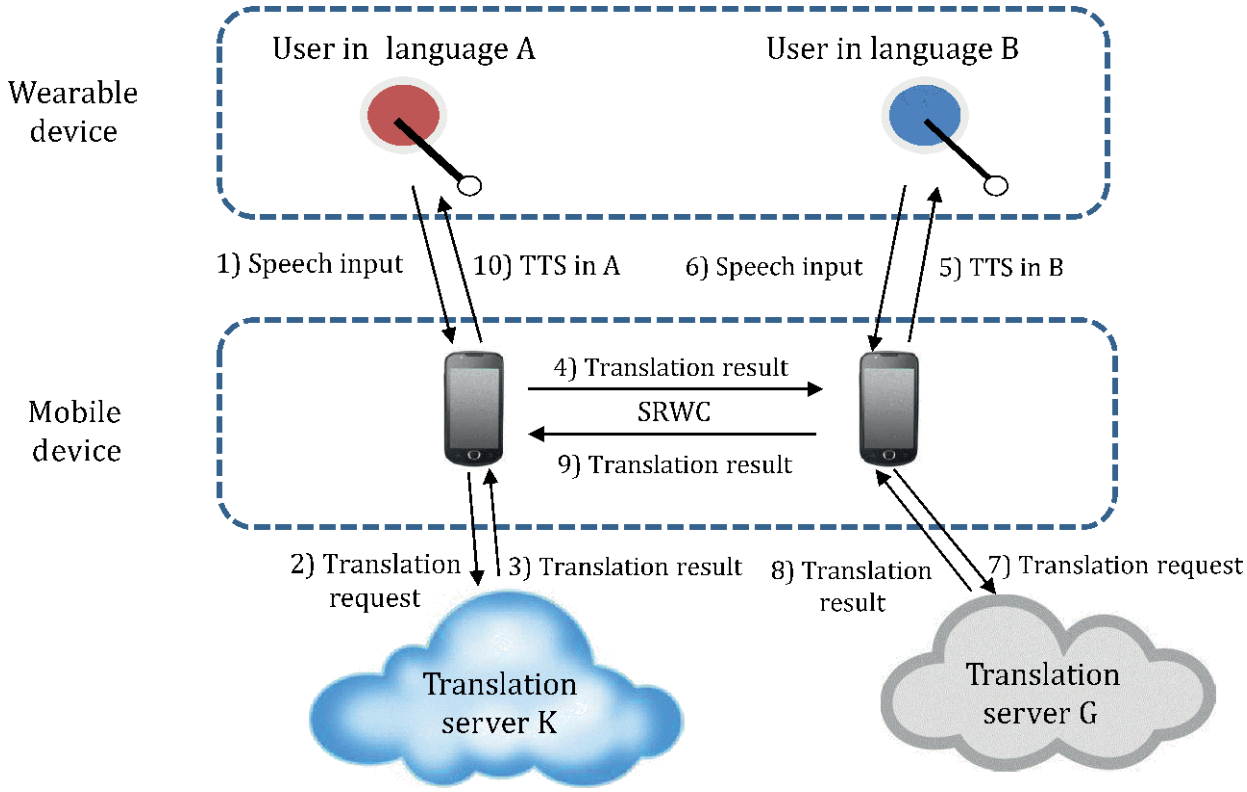


Figure 3 — The system architecture between two persons with embedded F2F speech translation devices

— In this configuration, the language A speech recognition module and the language A speech synthesizer are embedded in the mobile device of the user in language A, and the language B speech recognition module and the language B speech synthesizer are embedded in the mobile device of the user in language B.

— The A-to-B and B-to-A language translation modules reside in the translation server of translation service.

— The data format of (2), (3), (7) and (8) can be any format. For example, one can use Modality Conversion Markup Language[1].

— One of the mobile devices can be a fixed device with short range wireless communication capability. Tellers or box offices can use such an architecture.

The following steps are speech translation service steps between two persons with embedded F2F speech translation devices.

1) The user in language A speaks a sentence in language A. The language A speech recognition module embedded in the mobile device of the user recognizes the speech in language A, and outputs the corresponding text in language A.

2) The text in language A is translated into a text in language B of the same meaning through the A-to-B language translation module in the translation server K.

3) The translated text in language B is transferred to the mobile device of the user in language A,

4) And then transferred through short range wireless communication to the mobile device of the user in language B.

5) The language B speech synthesizer generates corresponding speech in language B.

6) Listening to the speech in language B, the user in language B answers in language B. The language B speech recognition module embedded in the mobile device of the user recognizes this speech in language B into text in language B. This recognized text is transferred to the B-to-A language translation module residing in the translation server G.

7) The text in language B is translated into a text in language A of the same meaning through the B-to-A language translation module residing in the translation server G.

8) The translated text in language A is transferred to the mobile device of the user in language B,

9) And then transferred to the mobile device of the user in language A through the short range wireless communication.

10) The language A speech synthesizer generates the corresponding speech in language A.

11) The steps (1) to (10) continues until both users accomplish their goals.

6.3 Two persons with remote speech translation functions

The system architecture between two persons with remote F2F speech translation devices is described in Figure 4.

Description of Figure 4 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (wearable device layer), the middle layer (mobile device layer), and the lower layer (speech recognition and translation server layer). This figure can also be divided into two vertical parts, i.e., the left part and the right part. The left part corresponds to the user in language A, and the right part corresponds to the user in language B.

The upper left part shows a wearable device of the user in language A. The middle left part shows a mobile device of the user in language A. There is an arrow from the wearable device to the mobile device with a text of “1) Speech input”. The lower left part shows a cloud shape with a text of “Speech recognition server (A)” and another cloud shape with a text of “Translation server K”. There is an arrow from the mobile device to the speech recognition server (A) with a text of “2) Speech recognition request”. There is another arrow from the speech recognition server (A) to the mobile device with a text of “3) Recognition result”. There is an arrow from the mobile device to the translation server K with a text of “4) Translation request”. There is another arrow from the translation server K to the mobile device with a text of “5) Translation result”.

The middle right part shows a mobile device of the user in language B. There is an arrow from the mobile device of the user in language A to the mobile device of the user in language B with a text of “6) Translation result”. The upper right part shows a wearable device of the user in language B. There is an arrow from the mobile device of the user in language B to the wearable device of the user in language B with a text of “7) TTS in B”. There is another arrow from the wearable device of the user in language B to the mobile device of the user in language B with a text of “8) Speech input”. The lower right part shows a cloud shape with a text of “Translation server G” and another cloud shape with a text of “Speech recognition server (B)”. There is an arrow from the mobile device of the user in language B to the speech recognition server (B) with a text of “9) Speech recognition request”. There is another arrow from the speech recognition server (B) to the mobile device of the user in language B with a text of “10) Recognition result”. There is an arrow from the mobile device of the user in language B to the translation server G with a text of “11) Translation request”. There is another arrow from the translation server G to the mobile device of the user in language B with a text of “12) Translation result”. There is an arrow from the mobile device of the user in language B to the mobile device of the user in language A with a text of “13) Translation result” and “SWRC”. There is an arrow from the mobile device of the user in language A to the wearable device of the user in language A with a text of “14) TTS in A”.

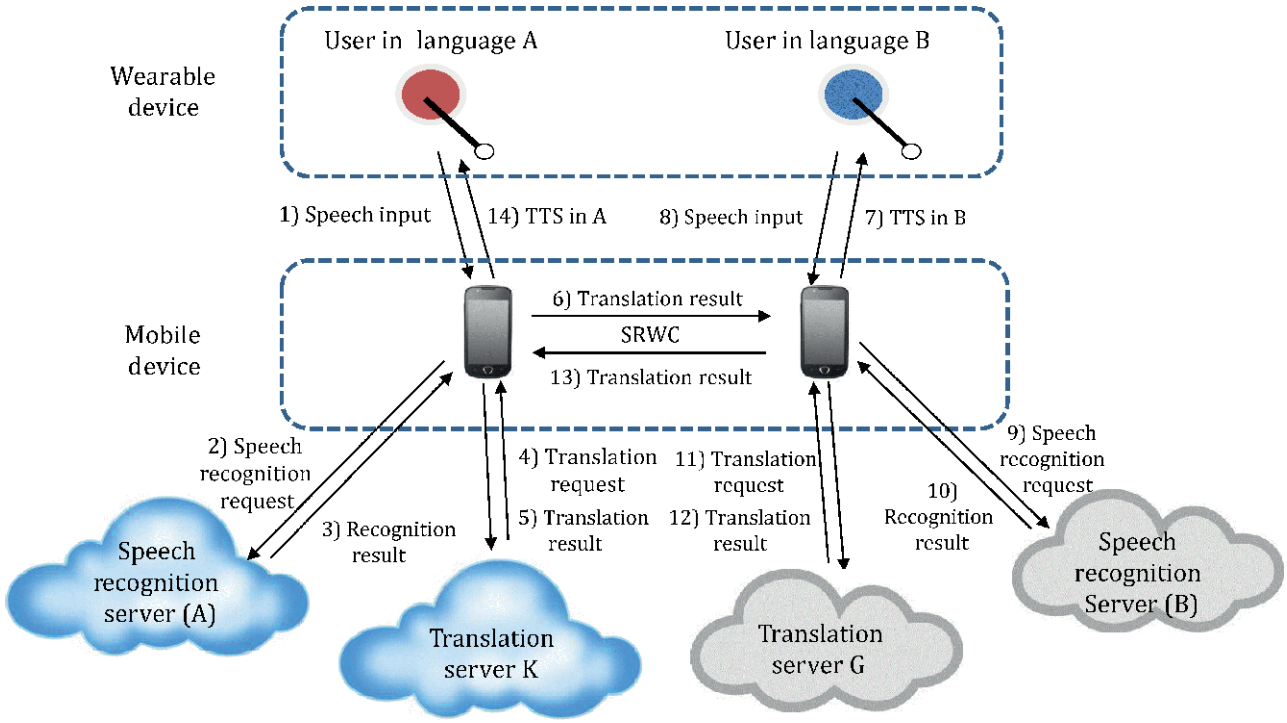


Figure 4 — The system architecture between two persons with remote F2F speech translation devices

— In this configuration, the language A speech synthesizer is embedded in the mobile device of the user in language A, and the language B speech synthesizer is embedded in the mobile device of the user in language B.

— The A-to-B and B-to-A language translation modules, the language A speech recognition module and the language B speech recognition module reside in remote environment.

— The speech synthesizer can also be in the remote environment.

— One of the mobile devices can be a fixed device with short range wireless communication capability. Tellers or box offices can use such architecture.

The following steps are speech translation service steps between two persons with remote F2F speech translation devices.

1) The user in language A speaks a sentence in language A.

2) The language A speech recognition module residing in the remote environment recognizes the speech in language A, and output corresponding text in language A.

3) The recognized text in language A is transferred to the mobile device of the user in language A.

4) The text in language A is translated into a text in language B of the same meaning through the A-to-B language translation module residing in the translation server K.

5) The translated text in language B is transferred to the mobile device of the user in language A,

6) And then transferred through short range wireless communication to the mobile device of the user in language B.

7) The language B speech synthesizer generates corresponding speech in language B.

8) Listening the speech in language B, the user in language B answers in language B.

9) The language B speech recognition module residing in the remote environment recognizes this speech in language B into text in language B.

10) The recognized text is transferred to the mobile device of the user in language B,

11) And then transferred to the B-to-A language translation module resided in the translation server G. The text in language B is translated into a text in language A of the same meaning through the B-to-A language translation module residing in the translation server G.

12) The translated text in language A is transferred to the mobile device of the user in language B,

13) And then transferred to the mobile device of the user in language A through the short range wireless communication.

14) The language A speech synthesizer generates corresponding speech in language A.

15) The steps (1) to (14) continues until both users accomplish their goals.

6.4 Mixture of 6.1 and 6.2

The system architecture of mixed environment is described in Figure 5.

Description of Figure 5 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (wearable device layer), the middle layer (mobile device layer), and the lower layer (speech recognition and translation server layer). This figure can also be divided into two vertical parts, i.e., the left part and the right part. The left part corresponds to the user in language A, and the right part corresponds to the user in language B.

The upper left part shows a wearable device of the user in language A. The middle left part shows a mobile device of the user in language A. There is an arrow from the wearable device to the mobile device with a text of “1) Speech input”. The lower left part shows a cloud shape with a text of “Speech recognition server (A)” and another cloud shape with a text of “Translation server K”. There is an arrow from the mobile device to the speech recognition server (A) with a text of “2) Speech recognition request”. There is another arrow from the speech recognition server (A) to the mobile device with a text of “3) Recognition result”. There is an arrow from the mobile device to the translation server K with a text of “4) Translation request”. There is another arrow from the translation server K to the mobile device with a text of “5) Translation result”.

The middle right part shows a mobile device of the user in language B. There is an arrow from the mobile device of the user in language A to the mobile device of the user in language B with a text of “6) Translation result”. The upper right part shows a wearable device of the user in language B. There is an arrow from the mobile device of the user in language B to the wearable device of the user in language B with a text of “7) TTS in B”. There is another arrow from the wearable device of the user in language B to the mobile device of the user in language B with a text of “8) Speech input”. The lower right part shows a cloud shape with a text of “Translation server G”. There is an arrow from the mobile device of the user in language B to the translation server G with a text of “9) Translation request”. There is another arrow from the translation server G to the mobile device of the user in language B with a text of “10) Translation result”. There is an arrow from the mobile device of the user in language B to the mobile device of the user in language A with a text of “11) Translation result” and “SWRC”. There is an arrow from the mobile device of the user in language A to the wearable device of the user in language A with a text of “12) TTS in A”.

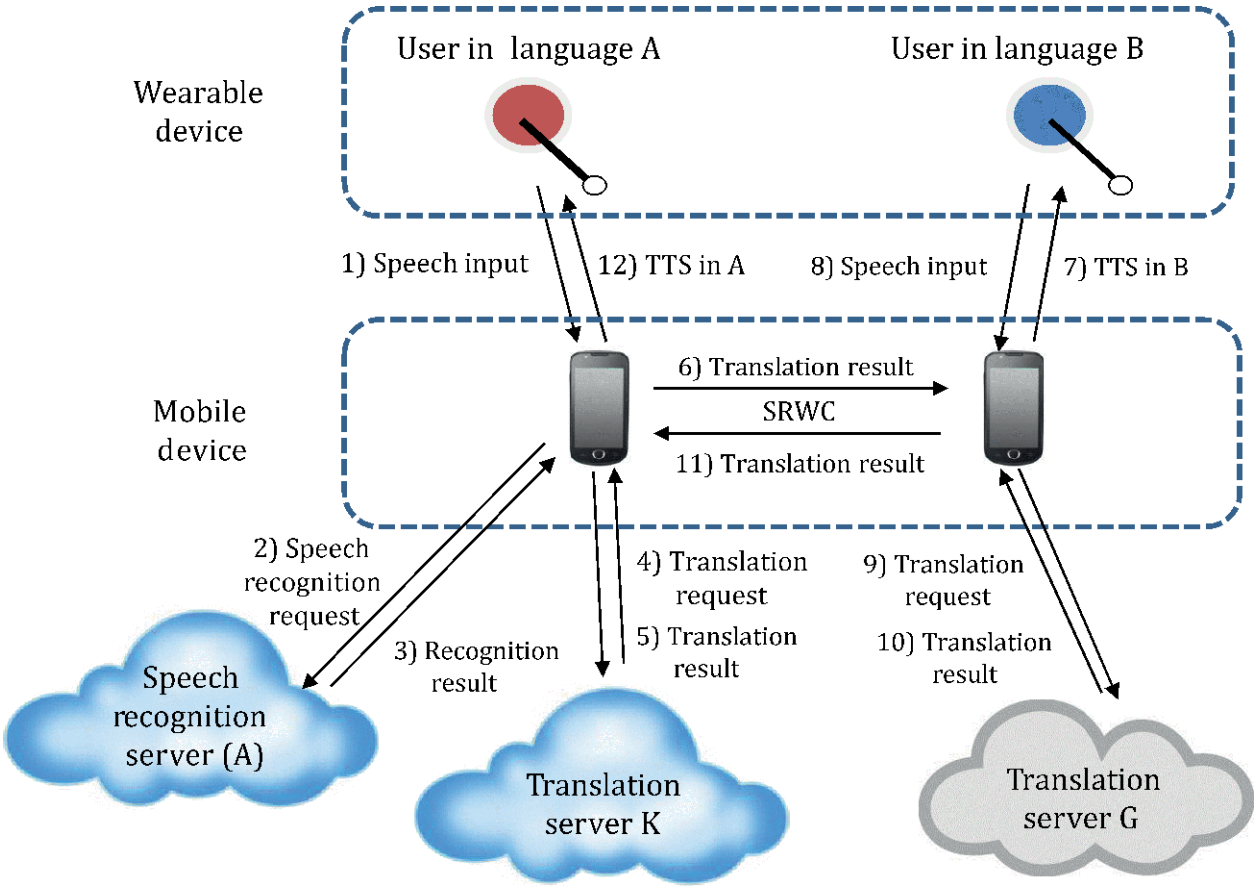


Figure 5 — The system architecture between two persons with remote F2F speech translation devices

— In this configuration, the language A speech synthesizer is embedded in the mobile device of the user in language A, the language B speech recognition module and the language B speech synthesizer are embedded in the mobile device of the user in language B.

— The A-to-B and B-to-A language translation modules, and the language A speech recognition module reside in remote environment.

The following steps are speech translation service steps between two persons with mixed F2F speech translation devices.

1) The user in language A speaks a sentence in language A.

2) The language A speech recognition module residing in a remote environment recognizes the speech in language A, and output corresponding text in language A.

3) The recognized text in language A is transferred to the mobile device of the user in language A.

4) The text in language A is translated into a text in language B of the same meaning through the A-to-B language translation module residing in the translation server K.

5) The translated text in language B is transferred to the mobile device of the user in language A,

6) And then transferred through short range wireless communication to the mobile device of the user in language B.

7) The language B speech synthesizer generates corresponding speech in language B.

8) Listening the speech in language B, the user in language B answers in language B. The language B speech recognition module embedded in the mobile device recognizes this speech in language B into a text in language B.

9) This recognized text is transferred to the B-to-A language translation module residing in the translation server G. The text in language B is translated into a text in language A of the same meaning through the B-to-A language translation module.

10) The translated text in language A is transferred to the mobile device of the user in language B,

11) And then transferred to the mobile device of the user in language A through the short range wireless communication.

12) The language A speech synthesizer generates corresponding speech in language A.

13) The steps (1) to (12) continues until both users accomplish their goals.

6.5 Adding one more speaker to F2F speech translation conversation

The system architecture among three persons with embedded F2F speech translation devices is described in Figure 6.

Description of Figure 6 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (wearable device layer), the middle layer (mobile device layer), and the lower layer (translation server layer). This figure can also be divided into three vertical parts, i.e., the left part, the center part, and the right part. The left part corresponds to the user in language A, the middle part corresponds to the user in language B, and the right part corresponds to the user in language C. This situation occurs when a new user in language C joins to the existing face-to-face speech translation conversation between the user in language A and the user in language B.

The upper left part shows a wearable device of the user in language A and the middle left part shows a mobile device of the user in language A. The upper center part shows a wearable device of the user in language B and the middle center part shows a mobile device of the user in language B. The upper right part shows a wearable device of the user in language C and the middle right part shows a mobile device of the user in language C.

There is an arrow from the wearable device to the mobile device of the user in language C with a text of “1) Speech input”. The lower left part shows a cloud shape with a text of “Translation server L” that indicates language translation server of the user in language C. There is an arrow from the mobile device to the translation server L with a text of “2) Translation request”. There is another arrow from the translation server L to the mobile device of the user in language C with a text of “3) Translation result (A, B)” which means that the sentence in language C has been translated into the sentences in language A and B.

There are two arrows from the mobile device of the user in language C to the mobile devices of the users in language A and B with texts of “4) Translation result (A)” and “4) Translation result (B)”. There is an arrow from the mobile device of the user in language A to the wearable device of the user in language A with a text of “5) TTS in A”. There is another arrow from the mobile device of the user in language B to the wearable device of the user in language B with a text of “5) TTS in B”.

There is an arrow from the wearable device of the user in language B to the mobile device of the user in language B with a text of “6) Speech input” which means the user in language B has answered to the speech of the user in language C. The lower center part shows a cloud shape with a text of “Translation server G” that indicates language translation server of the user in language B. There is an arrow from the mobile device of the user in language B to the translation server G with a text of “7) Translation request”. There is another arrow from the translation server G to the mobile device of the user in language B with a text of “8) Translation result (A, C)”. There are two arrows from the mobile device of the user in language B to the mobile devices of the users in language A and C with a texts of “9)

Translation result (A)" and "9) Translation result (C)". There is an arrow from the mobile device of the user in language A to the wearable device of the user in language A with a text of "10) TTS in A". There is another arrow from the mobile device of the user in language C to the wearable device of the user in language C with a text of "10) TTS in A".

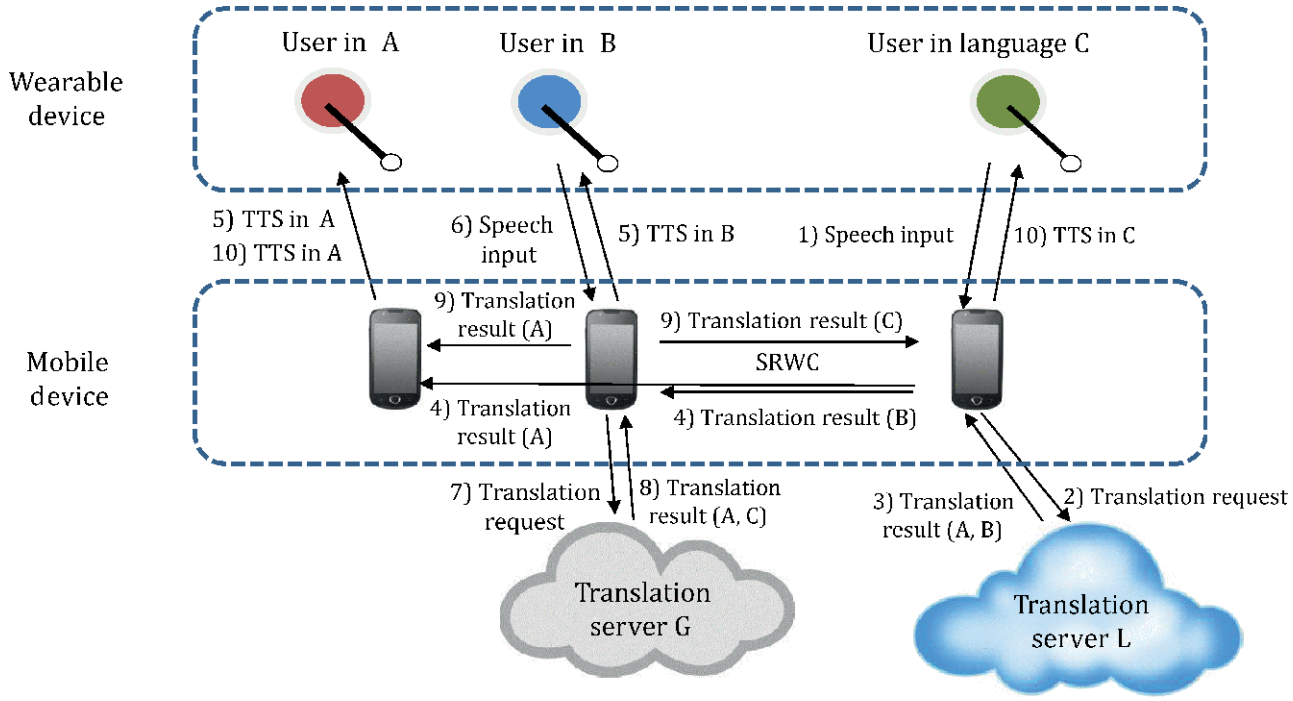


Figure 6 — The system architecture among three persons with embedded F2F speech translation devices

— This situation occurs when a new user in language C joins to the existing two persons' face-to-face speech translation conversation.

— In this configuration, all users have a speech recognition module and a speech synthesizer of his own language embedded in his mobile device.

— The language translation modules reside in the translation servers of translation service.

— Adding one more user to existing 3 or more persons' face-to-face speech translation conversation has the same steps of this architecture.

The following steps occur when a new user in language C joins to the existing two persons' face-to-face speech translation conversation.

1) The user in language C speaks a sentence in language C. The language C speech recognition module embedded in the mobile device of the user recognizes the speech in language C, and output corresponding text in language C.

2) The text in language C is translated into a text in languages A and B of the same meaning through the C-to-A and C-to-B language translation module residing in the translation server L.

3) The translated texts in language A and B are transferred to the mobile device of the user in language C.

4) The translated text in language A is then transferred through short range wireless communication to the mobile device of the user in language A. At the same time, the translated text in language B is transferred through the short range wireless communication to the mobile device of the user in language B. In this case, the information about the original speaker C shall be additionally transferred to the other users to increase naturalness of the synthesized speech.

5) The language A speech synthesizer generates corresponding speech in language A. At the same time, the language B speech synthesizer generates corresponding speech in language B.

6) Listening the speech in language B, the user in language B answers in language B. The language B speech recognition module embedded in the mobile device of the user recognizes this speech in language B into a text in language B. This recognized text is transferred to the B-to-A and B-to-C language translation module residing in the translation server L.

7) The text in language B is translated into a text in languages A and C of the same meaning through the B-to-A and B-to-C language translation module.

8) The translated texts in language A and C is transferred to the mobile device of the user in language B.

9) The translated text in language A is transferred to the mobile device of the user in language A through the short range wireless communication. At the same time, the translated text in language C is transferred to the mobile device of the user in language C through the short range wireless communication. In this case, the information about the original speaker B shall be additionally transferred to the other users to increase naturalness of the synthesized speech.

10) The language A speech synthesizer generates corresponding speech in language A. At the same time, the language C speech synthesizer generates corresponding speech in language C.

11) The steps (1) to (10) continues until all users accomplish their goals.

6.6 Two person with only one fixed F2F speech translation device

The system architecture between two persons with one fixed F2F speech translation device is described in Figure 7.

Description of Figure 7 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (user layer), the middle layer (fixed device layer), and the lower layer (remote server layer).

The upper left part shows a user in language A and the upper right part shows a user in language B. In the middle layer, there is only a fixed device such as a stand-alone computer. The lower left part shows a cloud shape with a text of "Translation server K" that indicates language translation server of the fixed device. The lower center part shows a cloud shape with a text of "Speech recognition server (B)" that changes speeches in language B into sentences in language B. The lower right part shows a cloud shape with a text of "Speech synthesizer (B)" that changes sentences in language B into speeches in language B.

There is an arrow from the user in language A to the fixed device with a text of "1) Speech input". There is an arrow from the fixed device to the translation server K with a text of "2) Translation request". There is another arrow from the translation server K to the fixed device with a text of "3) Translation result (B)" which means that the sentence in language A has been translated into the sentence in language B. There is an arrow from the fixed device to the speech synthesizer (B) with a text of "4) Speech synthesis request". There is another arrow from the speech synthesizer (B) to the fixed device with a text of "5) Synthesized speech". There is an arrow from the fixed device to the user in language B with a text of "6) Speech (B)".

There is an arrow from the user in language B to the fixed device B with a text of "7) Speech input" which means the user in language B has answered to the speech of the user in language A. There is an arrow from the fixed device to the speech recognition server (B) with a text of "8) Speech recognition request". There is another arrow from the speech recognition server (B) to the fixed device with a text of "9) Recognized result". There is an arrow from the fixed device to the translation server K with a text of "10) Translation request". There is another arrow from the translation server K to the fixed device with a text of "11) Translation result (A)". There is an arrow from the fixed device to the user in language A with a text of "12) TTS in A".

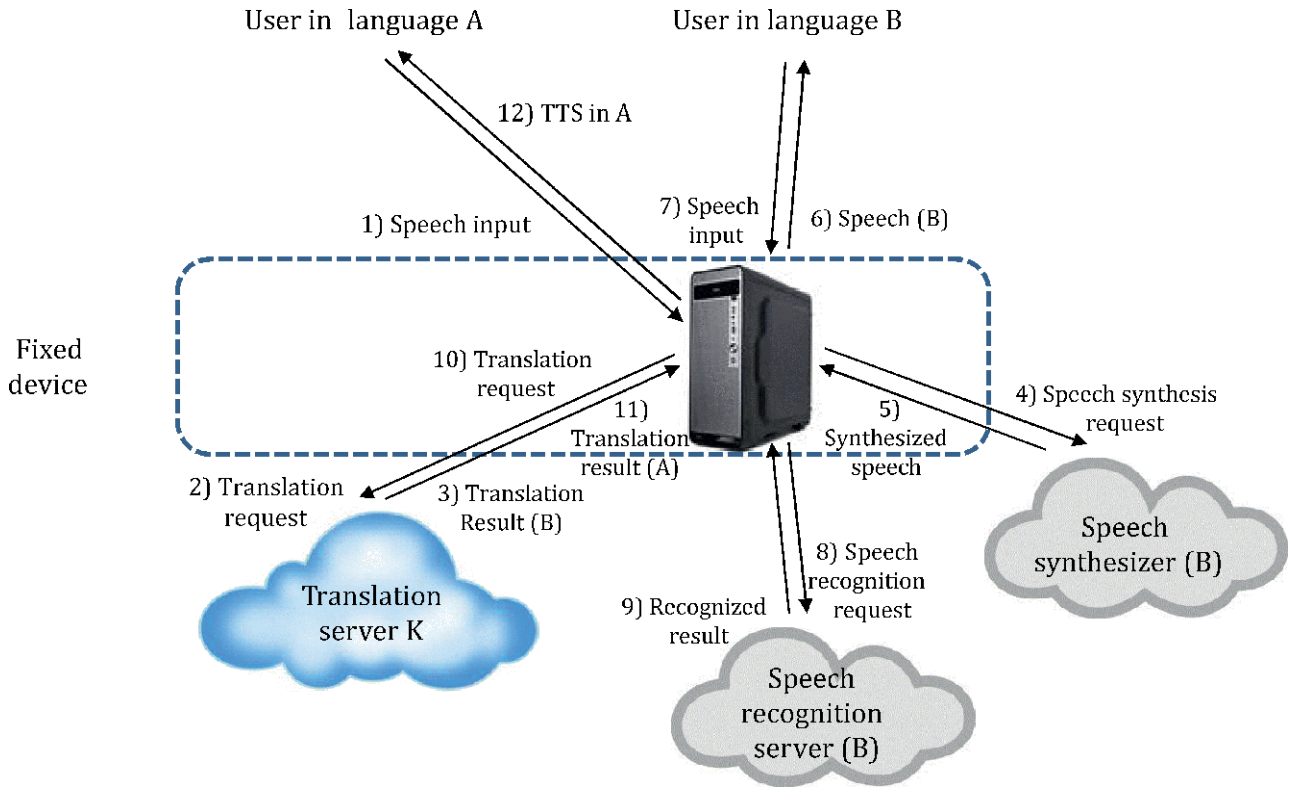


Figure 7 — The system architecture between two persons with one fixed F2F speech translation device

— In this configuration, the language A speech recognition module and the language A speech synthesizer are embedded in the fixed device. However, they can also be in remote environment.

— The language B speech recognition module and the language B speech synthesizer are resided in the remote environment. However, they can also be embedded in the fixed device.

— The A-to-B and B-to-A language translation modules reside in the translation server of translation service.

— In this configuration, the users should control UI set-up of Figure 1 before speech translation service. Both users shall select their language and gender manually.

— Both users may use one microphone alternately. The user should select his language first and then speak a sentence so that the proper speech recognizer can recognize it. If two microphones are available, each user may use his own microphone. In this case, one does not need to select his language every time before his speech.

— A mobile device can be used instead of the fixed device.

The following steps are speech translation service steps between two persons with one fixed F2F speech translation device.

1) The user in language A speaks a sentence in language A. The language A speech recognition module embedded in the fixed device recognizes the speech in language A, and output corresponding text in language A.

2) The text in language A is translated into a text in language B of the same meaning through the A-to-B language translation module residing in the translation server K.

3) The translated text in language B is transferred to the fixed device.

4) The fixed device requests speech synthesis of the text in language B to the speech synthesizer (B) resided in remote environment.

5) The synthesized speech in language B is transferred to the fixed device.

6) The fixed device plays the synthesized speech in language B.

7) Listening to the speech in language B, the user in language B answers in language B to the microphone of the fixed device.

8) The fixed device requests speech recognition of this speech in language B to the speech recognition server (B) in the remote environment.

9) The speech recognition server (B) returns the recognized text in language B to the fixed device.

10) This recognized text in language B is transferred to the B-to-A language translation module residing in the translation server G.

11) The text in language B is translated into a text in language A of the same meaning through the B-to-A language translation module. The translated text in language A is transferred to the fixed device.

12) The language A speech synthesizer generates corresponding speech in language. The user in language A can hear the synthesized speech in language A through the speaker of the fixed device.

13) The steps (1) to (12) continues until both users accomplish their goals.

Annex A
(informative)

History of F2F speech translation

A.1 Introduction

Speech translation is an essential function to communicate with persons with different languages. To accomplish speech translation, one needs to have three functional components; i.e., speech recognition, language translation, and speech synthesis, as shown in Figure A1.

Description of Figure A1 for accessibility purpose:

This figure consists of three horizontal layers, i.e., the upper layer (Chinese to English layer), the middle layer (user layer), and the lower layer (English to Chinese layer). This middle layer can be divided into two vertical parts, i.e., the left part (the Chinese user) and the right part (the English user).

The upper layer has three boxes with two left-to-right arrows between the boxes. The three boxes have texts “Speech(CHN) Recognition”, “Language Translation”, and “Speech(ENG) Synthesis”, respectively, from left to right.

The middle left part shows a Chinese user with a Chinese text “有BB霜吗?” on the top and another Chinese text “要哪个牌子的?” underneath. The middle right part shows a American user with a English text “Do you have BB cream?” on the top and another English text “Which brand do you prefer?” underneath.

The lower layer has three boxes with two right-to-left arrows between the boxes. From right to left, the three boxes have texts “Speech(ENG) Recognition”, “Language Translation”, and “Speech(CHN) Synthesis”, respectively.

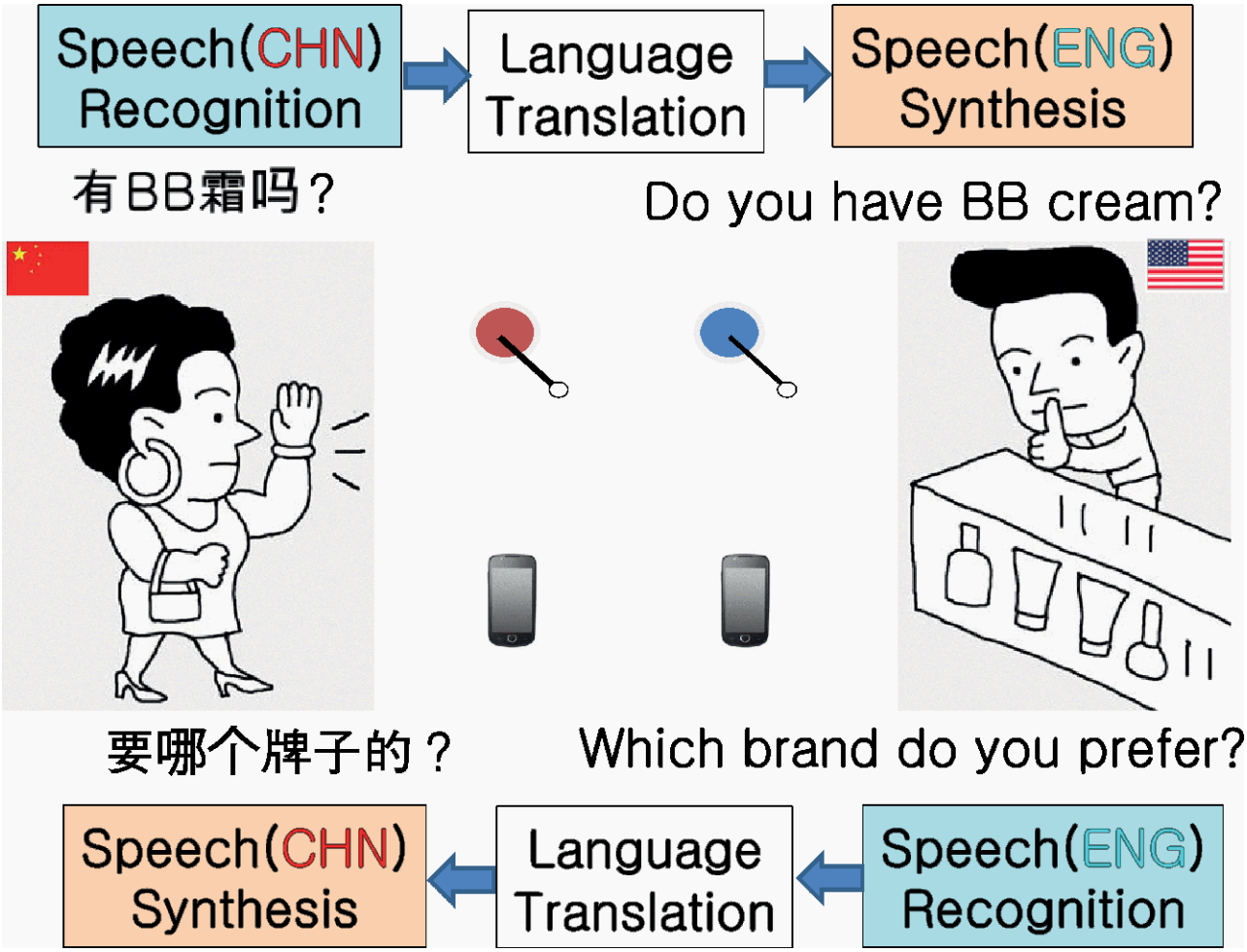


Figure A1 — Functional components of F2F speech translation

A Chinese woman approaches to an American cosmetics store, and asks for BB cream in Chinese language. The Chinese speech recognition module should recognize the Chinese speech into Chinese sentence. The Chinese-to-English language translation module translates the Chinese sentence into English sentence with the same meaning. The English speech synthesis module generates English speech from the English sentence. The American salesman asks the brand preference in English. The English speech recognition module should recognize the English speech into English sentence. The English-to-Chinese language translation module translates the English sentence into Chinese sentence with the same meaning. The Chinese speech synthesis module generates Chinese speech from the Chinese sentence.

In the previous days, one hires a human translator with multi-lingual capability to accomplish this goal. In the beginning of 21st century, automatic speech recognition, automatic language translation, and speech synthesis technologies have been developed, and the quality of these technologies is enough to serve as components of speech translation. Various speech translation services as well as application software for mobile devices appear in the market. However, to employ speech translation services, one needs to learn how to use the services and/or application software. This user interface method can be called “the 2nd generation translation” while “the 1st generation translation” means hiring human translators.

The goal of this standard is to establish a user interface method in F2F speech translation that has no pre-learning. In everyday life, people communicate with other people of same language easily. The user interface of this standard is just the same as that of communication among the people of same language. This method can be called the “3rd generation” or zero-effort speech translation.

In this annex, comparison among the three generations of F2F speech translation is given in the point of system functions and performances.

A.2 1st generation: human translator

Until 20th century, one used to hire a human translator to visit a place with different languages if he cannot speak that language. The human translator in figure A2 tries his best to help the employer to do whatever the employer wants to do.

Description of Figure A2 for accessibility purpose:

This figure consists of three vertical parts, i.e., the left part (Chinese user), the middle layer (human translator), and the right part (English user). The left part shows a Chinese user with a Chinese text "有BB霜吗?" on the top. The right part shows an American user with an English text "Which brand do you prefer?" underneath. In the middle, there is a human translator with an English text "Do you have BB cream?" on the top and another Chinese text "要哪个牌子的?" underneath.

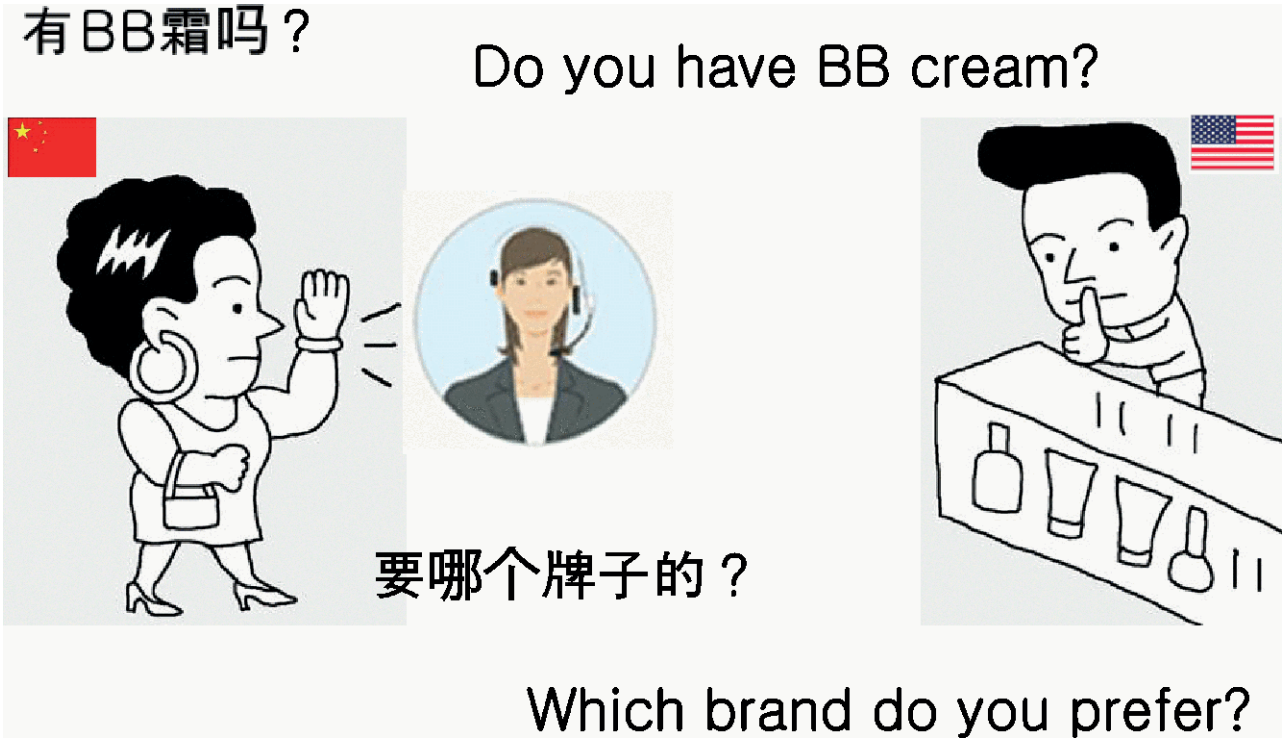


Figure A2 — 1st generation F2F speech translation: human translator

It is very easy for the employer to translate his speech because he just needs to speak to the human translator what he wants to do. It takes time to communicate since the human translator needs to translate the sentence and talks to the domestic person. If the domestic person answers, the human translator also translates it, and talks to the employer. Thus, it takes time to communicate with a person with different language.

The quality of 1st generation speech translation in this case directly depends on the human translator. The quality of speech translation strongly depends on the salary of the human translator, which is usually expensive. The privacy of the conversation cannot be guaranteed, since human translators understand every bit of the conversation.

A.3 2nd generation: automatic speech translation

In the beginning of 21st century automatic speech recognition, automatic language translation, and speech synthesis have been developed, and the quality of these technologies are enough to serve as components of speech translation. Various speech translation services as well as application software for mobile devices appeared in the market. However, to employ speech translation services, one need to learn how to use the services and/or application software as shown in Figure A3.

Description of Figure A3 for accessibility purpose:

This figure consists of two horizontal layers, i.e., the upper layer (one text line) and the lower layer (user layer). This lower layer can be divided into the left part (the Chinese user) and the right part (the English user). The upper part is a text "1) Phone call to the salesman 2) select target language". The lower left part shows a Chinese user with a Chinese text "有BB霜吗?" on the top. The Chinese user carries a mobile device in her hand. There is a text "Do you have BB cream?" nearby the mobile device. The lower right part shows an American user with an English text "Which brand do you prefer?" underneath. The American user carries a mobile device. There is a Chinese text "要哪个牌子的?" nearby the mobile device.

Figure A3 — 2nd generation F2F speech translation: pre-trained speech translation

The user should know how to make a channel for speech translation such as the other's phone number or ID number. Even after the channel set-up, he should know how to start speech input to the speech translation system. The other domestic person should also know how to use the speech translation system well. If all this conditions meet, the speech translation can be done in near real time, since the current speed of speech translation is near real time.

The quality of automatic speech translation depends on the quality of speech translation services. The state-of-the-art sentence translation accuracy is 80 to 90% which is very helpful to the users. The cost of 2nd generation speech translation is much cheaper than the 1st generation speech translation. This speech translation scheme keeps good privacy, since no other person knows about the conversation between the two persons.

The most critical reason that prevents people from using speech translation technologies is the fact that one needs to learn how to use the services and/or application software. There is another critical reason that prevents people from using speech translation technologies. When using the 2nd generation face-to-face speech translation, one should use both hands, i.e., one for holding the mobile device, and the other for operation. If one holds luggage in his both hands, it is impossible to use 2nd generation speech translation.

A.4 3rd generation: zero-effort speech translation

The 3rd generation speech translation is a convenient user interface method in F2F speech translation that has no additional learning. In everyday life, people communicate with other people of the same language easily. The user interface of this standard is just the same as that of communication among the people of same language. Thus this method can be called zero-effort speech translation as shown in Figure A4.

Description of Figure A4 for accessibility purpose:

This figure consists of two parts, i.e., the left part (the Chinese user, a woman) and the right part (the English user, a man). The left part shows a Chinese user with a Chinese text “有BB霜吗?” on the top. The right part shows an American user with a wearable device in his ear. There is an English text “Do you have BB cream?” in a speech balloon at the wearable device of the American user. There is an English text “Which brand do you prefer?” underneath the American user. The Chinese user carries a wearable device in her ear. There is a Chinese text “要哪个牌子的?” in a speech balloon at the wearable device of the Chinese user.

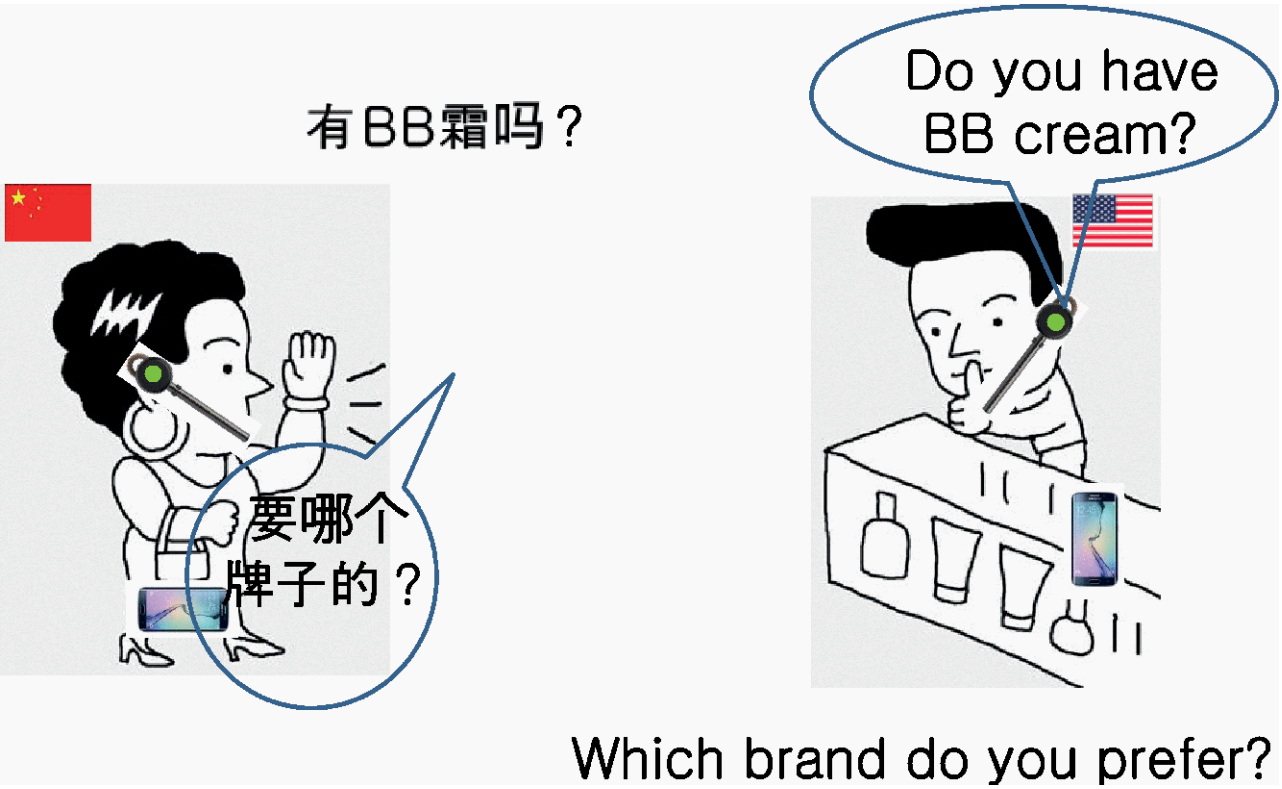


Figure A4 — 3rd generation F2F speech translation: zero-effort speech translation

The conversation starts when a foreign person with the 3rd generation speech translation devices approaches to the domestic person with the 3rd generation speech translation devices. (Detecting approaching action, the devices automatically starts communication to set up the channel and to select the target language.) The foreign person talks to the domestic person with his own language. The foreign speech is recognized, translated into domestic language, and synthesized into domestic speech.

In the foreign person's point of view, he has just approached to the domestic person, and talked to him with his own language. This is just the same situation that he starts conversation with the person in his country. He does not need to have any prior knowledge about how to use the speech translation system or services. When they finish conversation, anyone may leave the place. That automatically ends the conversation by detecting that two devices move away from each other.

The quality and cost of zero-effort speech translation is the same as those of automatic speech translation. This speech translation scheme also keeps good privacy. However, one does not need any pre-training to use the 3rd generation F2F speech translation services. The 3rd generation F2F speech translation allows hands-free operation as the conversation among people with same languages does. In this sense, zero-effort speech translation can be an ultimate method of face-to-face speech translation, which is described in this standard.

A.5 Summary

The table below summaries the comparison among the three methods in the sense of usability, system delay, quality, cost and privacy.

Speech translation methods	1st generation Human translator	2nd generation Automatic speech translation	3rd generation Zero-effort speech translation
Usability	Order to the human translator	Need to learn how to use Need both hands to use	Zero-effort Hands-free Same as domestic conversation
System delay	Long	Near real time	Near real time
Translation quality	Depends on human translator	80 ~ 90 % sentence accuracy	80 ~ 90 % sentence accuracy
Cost	Expensive	Cheap	Cheap
Privacy	Bad	Good	Good

Annex B
(informative)

An example scenario of F2F speech translation protocol

This Annex shows an example scenario of face-to-face speech translation between two persons.

A. A Chinese traveler approaches to an American salesman. The connection for face-to-face speech translation is automatically established.
B. The Chinese traveler says “有BB霜吗?”
C. The American salesman hears “Do you have BB cream?”
D. The American salesman asks “Which brand do you prefer?”
E. The Chinese traveler hears “要哪个牌子的?”
F. The Chinese traveler answers “C公司是多少?.”
G. The American salesman hears “How much is C company’s?”
H. The American salesman answers “25 dollars.”
I. The Chinese traveler hears “25美元. ”
J. The Chinese traveler pays 50 dollars and gets 2 BB creams. The Chinese traveler leaves the cosmetics store. The face-to-face speech translation is automatically disconnected.

Bibliography

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