INTERNATIONAL STANDARD

ISO/IEC 20382-2

First edition
2017-10

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

Reference number
ISO/IEC 20382-2:2017(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

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form a 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 organizations 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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

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. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 35, *User interfaces*.

A list of all parts in the ISO/IEC 20382- series can be found on the ISO website.

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 global environments. Automatic speech translation systems have existed for a long time, but they have functional limitations as well as technical ones with regard to usability and accessibility. Annex A shows a history of face-to-face speech translation.

One reason for these limitations is the diversity of the languages currently used. It is difficult to support many languages by one or several speech translation systems. A flexible and interoperable standardized framework is needed to work with all different languages utilizing many speech translation systems already developed in many countries. Other considerations to make a natural and usable speech translation service possible include applying users' characteristics within 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 document aims to enable face-to-face speech translation among people with different languages. The three technologies, i.e., speech recognition, language translation, and speech synthesis technologies, are mature enough to build a speech translation function. There are many face-to-face speech translation devices and/or services using mobile devices. However, the user needs to learn how to use the service and needs to use both hands to control the speech translation system. If the user wishes to use only one hand, which is usually the case, he or she cannot use the current speech translation systems and/or services. To overcome this usability issue, this document suggests a method that exactly follows the conversation among people with the same language. The method in this document is hands-free, and does not require any pre-training. In this sense, this method is the 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 the 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, thus providing a framework to support a convenient speech translation service in face-to-face situations. This document is applicable to speech translation devices, servers and communication protocols among speech translation servers and clients in a high-level approach. This document also defines various system architectures in different environments. This document is not applicable to defining speech recognition engines, language translation engines and speech synthesis engines.

2 Normative references

There are no normative references in this document.

3 Terms, definitions and abbreviated terms

3.1 Terms and definitions

No terms and definitions are listed in this document.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

3.2 Abbreviated terms

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

4 Overview of face-to-face speech translation

4.1 General

A face-to-face (F2F) speech translation system enables users of different languages in a face-to-face situation to communicate with each other in spoken languages by providing machine-generated translation results. A 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.

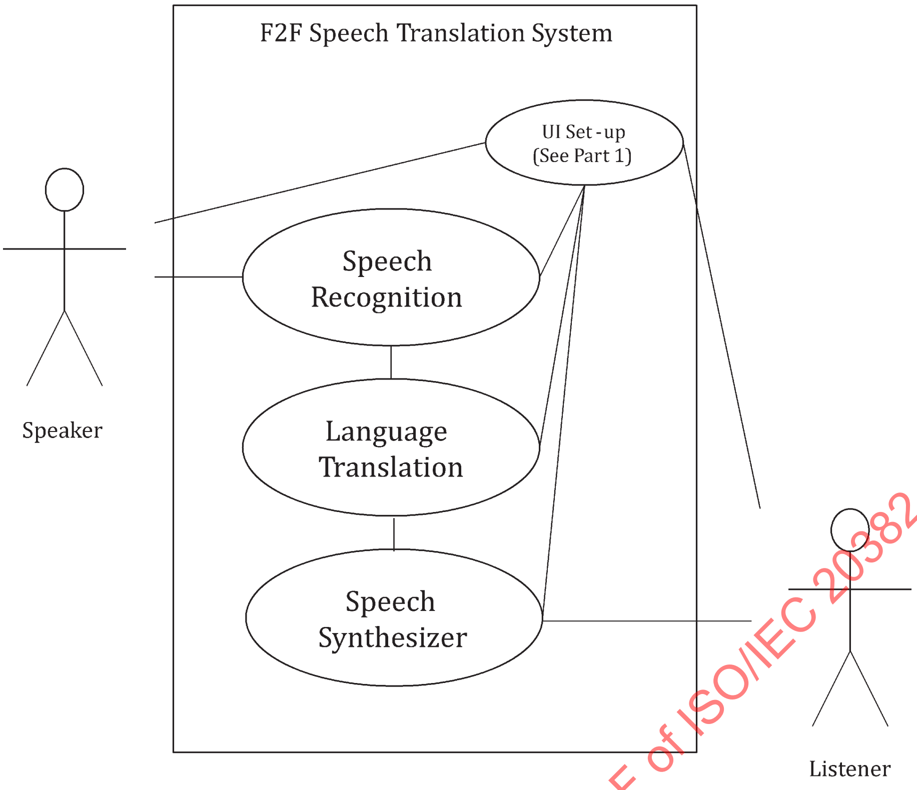


Figure 1 — Functional components of F2F speech translation

4.2 Functional components of F2F speech translation

For F2F speech translation, the speaker and the listener shall set up a UI (see ISO/IEC 20382-1.).

The functions of each component in Figure 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 with the same meaning through the language translation module.

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

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

6) Steps (2) to (5) continue until the users accomplish their goals.

5 Functional requirements

5.1 General requirement

Provides general requirements regarding face-to-face speech translation:

- there are three remote services in this document, 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 naturally as in everyday conversation;
- the translation system should allow the users to start a translation session as quickly 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 users to have a session with multiple users;
- the translation system should allow the users to add additional participants after the session has started.

5.2 Speech recognition requirements

Provides the requirements regarding the speech recognition module of face-to-face speech translation:

- the speech recognition module shall recognize the speech and provide it in 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 the MIME format;
- the output of the speech recognition module should be written in utf-8 format (see IETF RFC 2279 (1998)).

NOTE This document does not specify the data format of the speech nor that of the text since there are many off-the-shelf speech recognition modules with various input and output data formats.

5.3 Language translation requirements

Provides requirements for the user language translation module of face-to-face speech translation:

- the language translation module shall translate text from a source language into 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 an intermediate language to accomplish the language translation. One should translate the source language to the intermediate language, and then the intermediate language to the 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 from languages with the same word order as the source language or the target language.

NOTE This document does not specify the data formats of the input and output texts since there are many off-the-shelf language translation modules with various input and output data formats.

5.4 Speech synthesizer requirements

Provides requirements for the speech synthesizer of face-to-face speech translation:

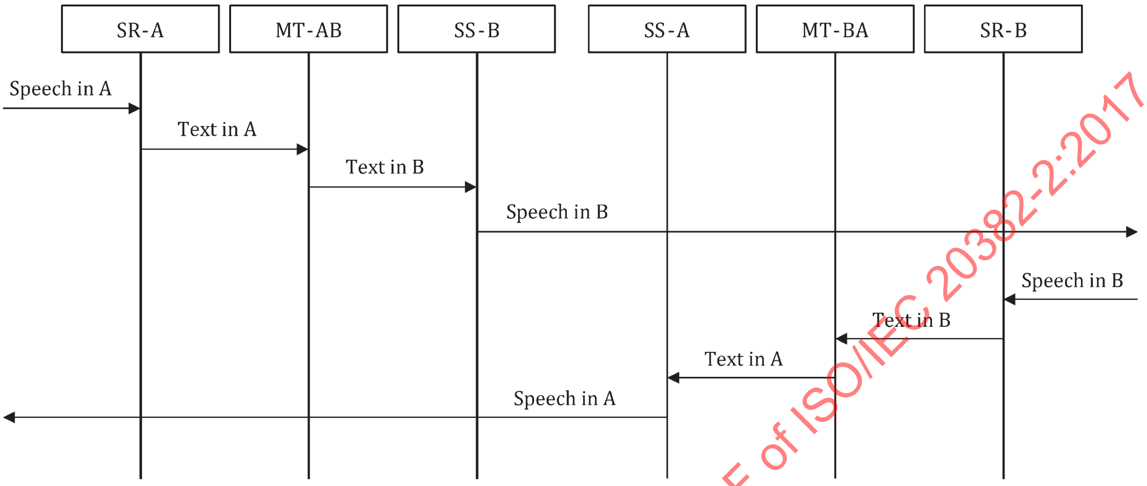
- the speech synthesizer shall generate the corresponding speech from text of the same language;
- in face-to-face speech translation the synthesized speech should be as close as possible to that of the original speaker to increase the natural feel of the conversation. The gender of the synthesized speech in language B should be the same as that of the user in language A. The natural feeling can be increased if the base frequency, speed, prosody and/or speech colour 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 (see IETF RFC 2279 (1998)).

NOTE This document does not specify the data format of the speech nor that of the text since there are many off-the-shelf speech synthesizers 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.



- 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 — Sequence diagram

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.

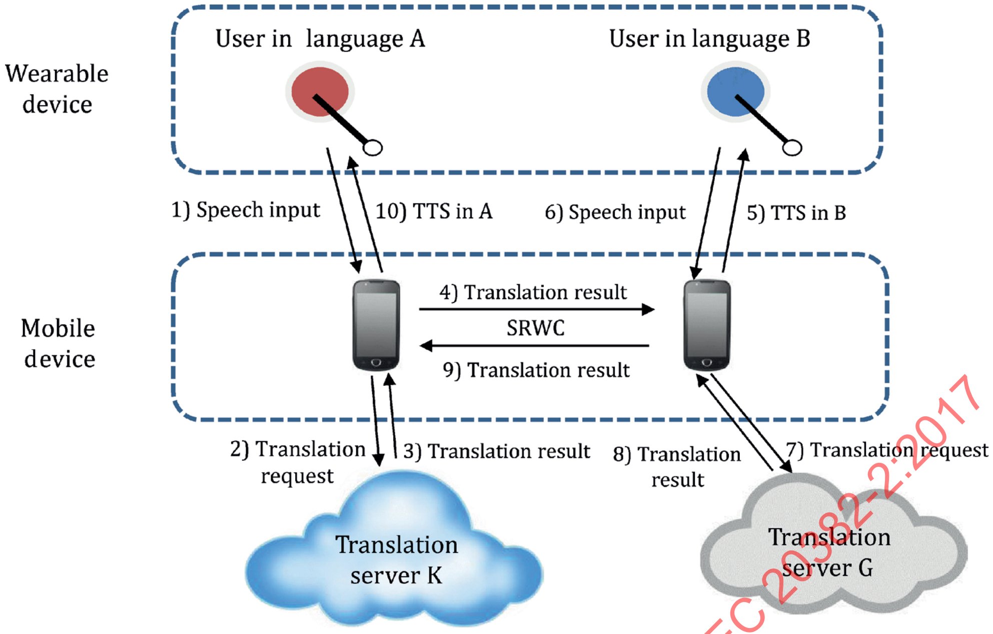


Figure 3 — System architecture between two persons with embedded F2F speech translation devices

— 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 the translation service.

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

— 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. Annex B shows an example scenario of face-to-face speech translation protocol.

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 text in language B with the same meaning through the A-to-B language translation module in translation server K.

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

4) The translated text in language B is then transferred through short range wireless communication to the mobile device of the user in language B.

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

6) After 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 and outputs the corresponding text in language B. This recognized text is transferred to the B-to-A language translation module residing in translation server G.

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

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

9) The translated text in language A is 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) 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.

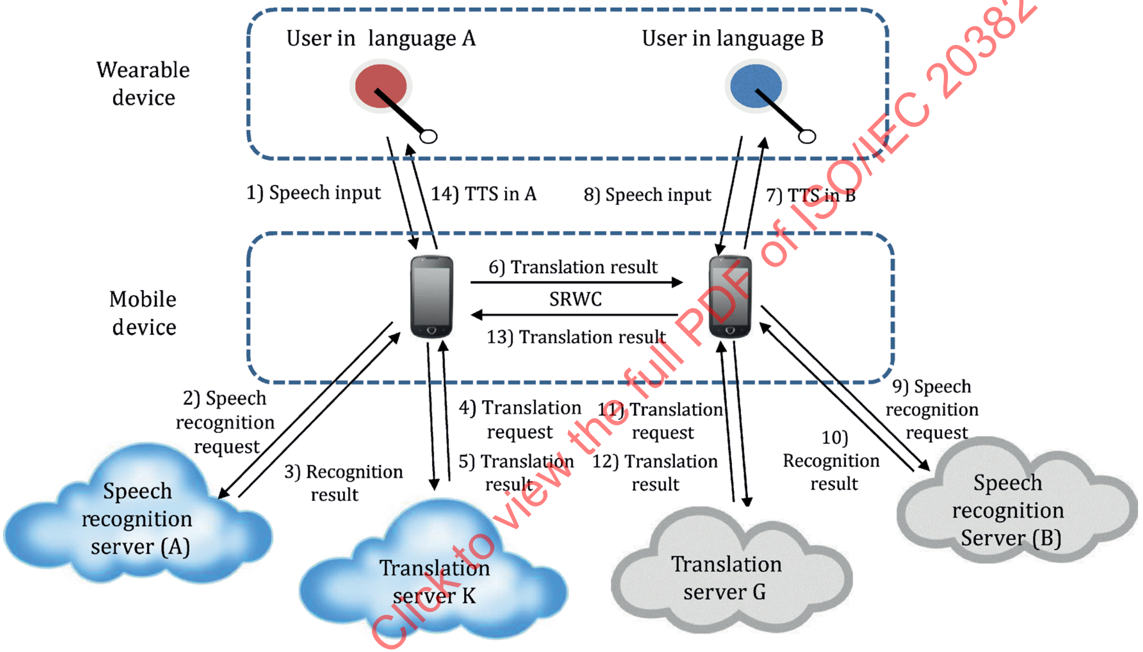


Figure 4 — System architecture between two persons with remote F2F speech translation devices

- 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 a remote environment.
- The speech synthesizer can also reside 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 outputs 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 text in language B with the same meaning through the A-to-B language translation module residing in translation server K.

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

6) The translated text in language B is 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) After listening to 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 and outputs corresponding text in language B.

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

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

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

13) The translated text in language A is then transferred to the mobile device of the user in language A through short range wireless communication.

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

15) Steps (1) to (14) continue until both users accomplish their goals.

6.4 Mixture of 6.2 and 6.3

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

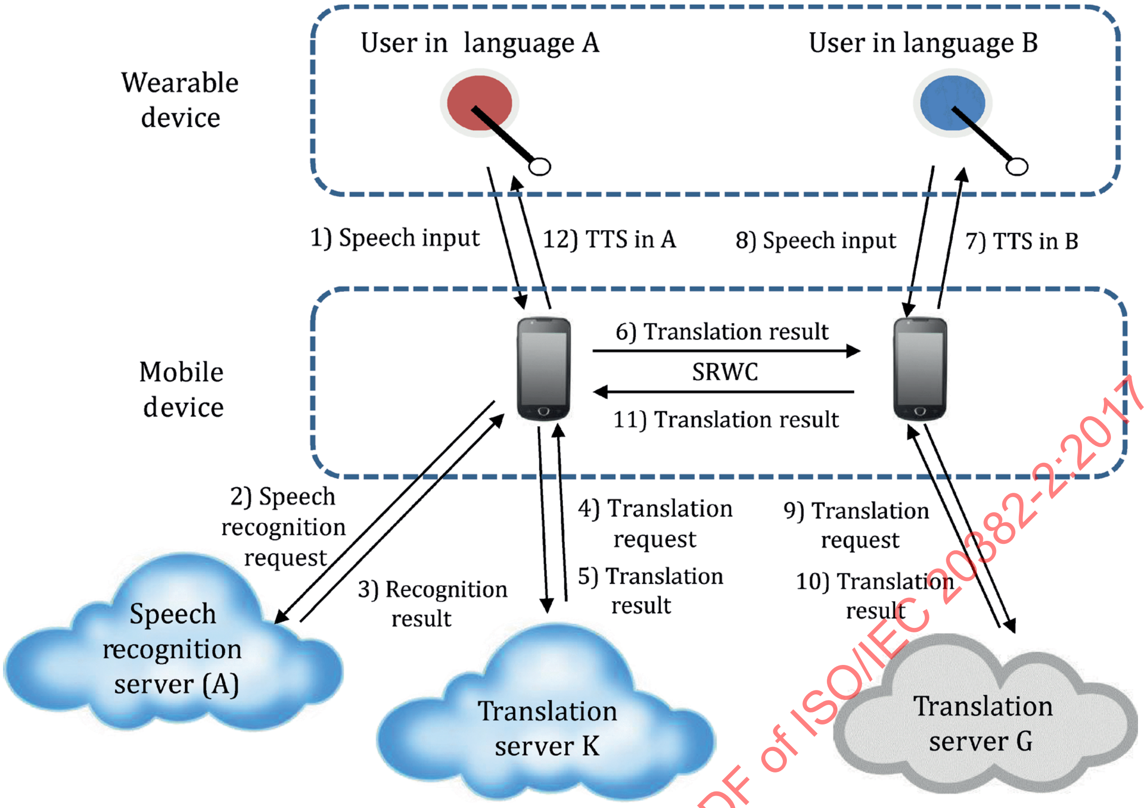


Figure 5 — System architecture between two persons with remote F2F speech translation devices

- 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 a 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 outputs 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 text in language B with the same meaning through the A-to-B language translation module residing in translation server K.

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

6) The translated text in language B is 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) After 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 recognizes this speech in language B and outputs corresponding text in language B.

9) This recognized text is transferred to the B-to-A language translation module residing in translation server G. The text in language B is translated into text in language A with 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) The translated text in language A is then transferred to the mobile device of the user in language A through short range wireless communication.

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

13) Steps (1) to (12) continue 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.

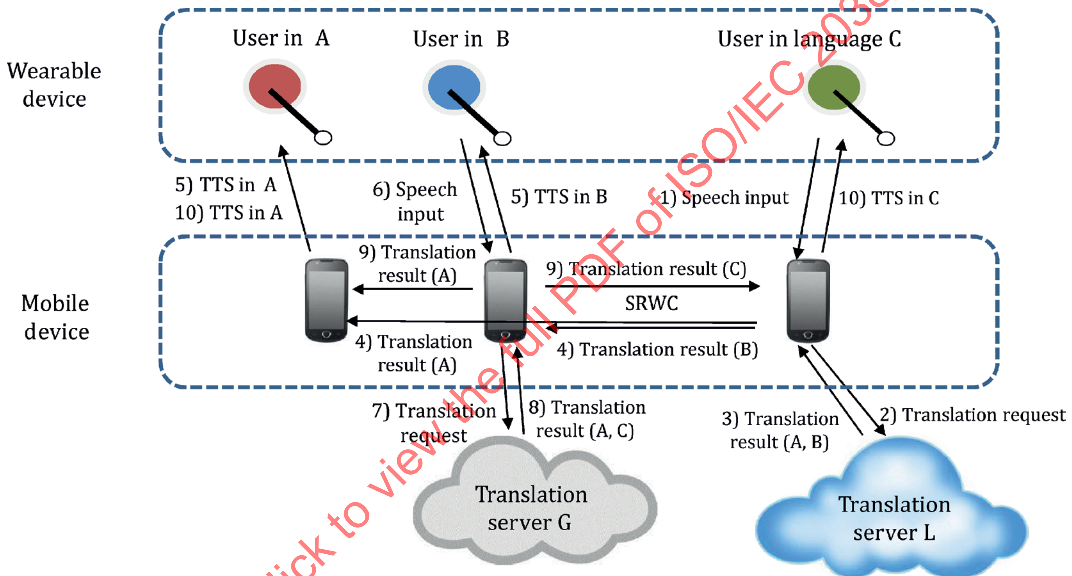


Figure 6 — System architecture among three persons with embedded F2F speech translation devices

- This situation occurs when a new user in language C joins 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 their own language embedded in a mobile device.
- The language translation modules reside in the translation servers of the translation service.
- Adding one more user to the existing three or more persons' face-to-face speech translation conversation uses the same steps in this architecture.

The following steps occur when a new user in language C joins 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 outputs corresponding text in language C.

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

3) The translated texts in languages 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 the natural feeling 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) After 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 and outputs corresponding text in language B. This recognized text is transferred to the B-to-A and B-to-C language translation module residing in translation server L.

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

8) The translated texts in languages A and C are 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 the natural feeling 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) 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.

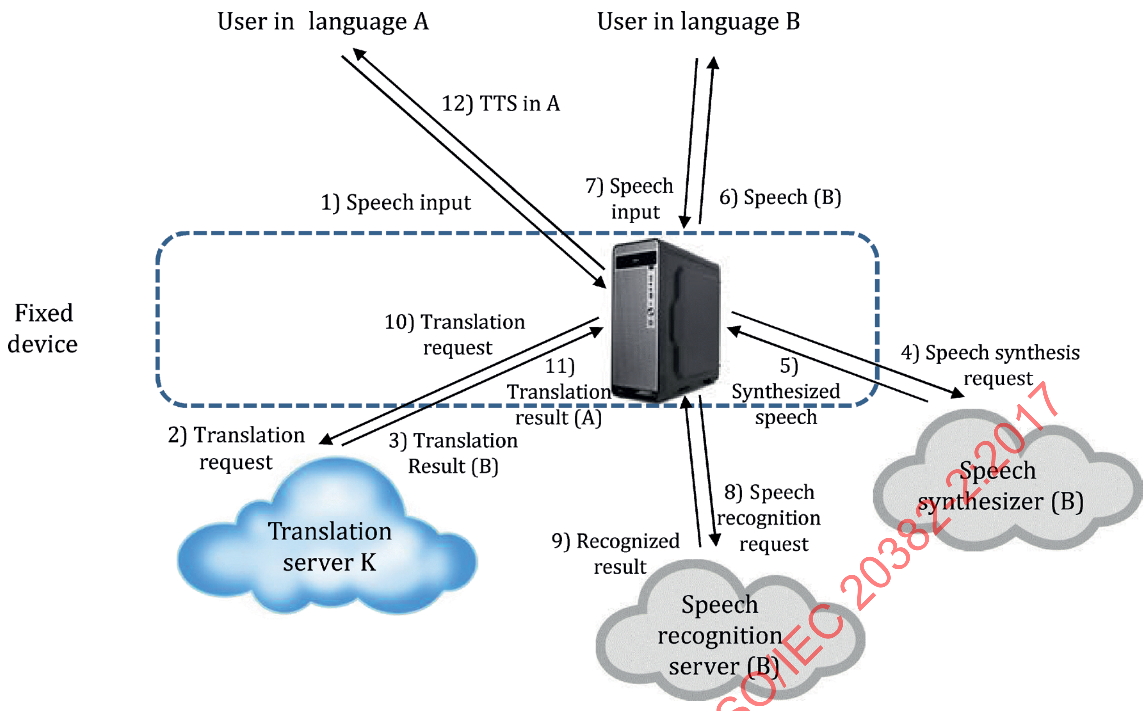


Figure 7 — System architecture between two persons with one fixed F2F speech translation device

- 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 placed in a remote environment.

- The language B speech recognition module and the language B speech synthesizer reside 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 the translation service.

- In this configuration, the users should control the UI set-up of Figure 1 before using the 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 identify it. If two microphones are available, each user may use his own microphone. In this case, they do not need to select their language every time before speaking.

- 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 outputs corresponding text in language A.

2) The text in language A is translated into text in language B with the same meaning through the A-to-B language translation module residing in 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) residing in the 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) After listening to the speech in language B, the user in language B answers in language B using 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 translation server G.

11) The text in language B is translated into text in language A with 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 A. The user in language A can hear the synthesized speech in language A through the speaker of the fixed device.

13) Steps (1) to (12) continue until both users accomplish their goals.

Annex A
(informative)

History of F2F speech translation

A.1 General

Speech translation is an essential function to communicate with persons with different languages. To accomplish speech translation, one needs to have three functional components as shown in Figure A.1: speech recognition, language translation, and speech synthesis.

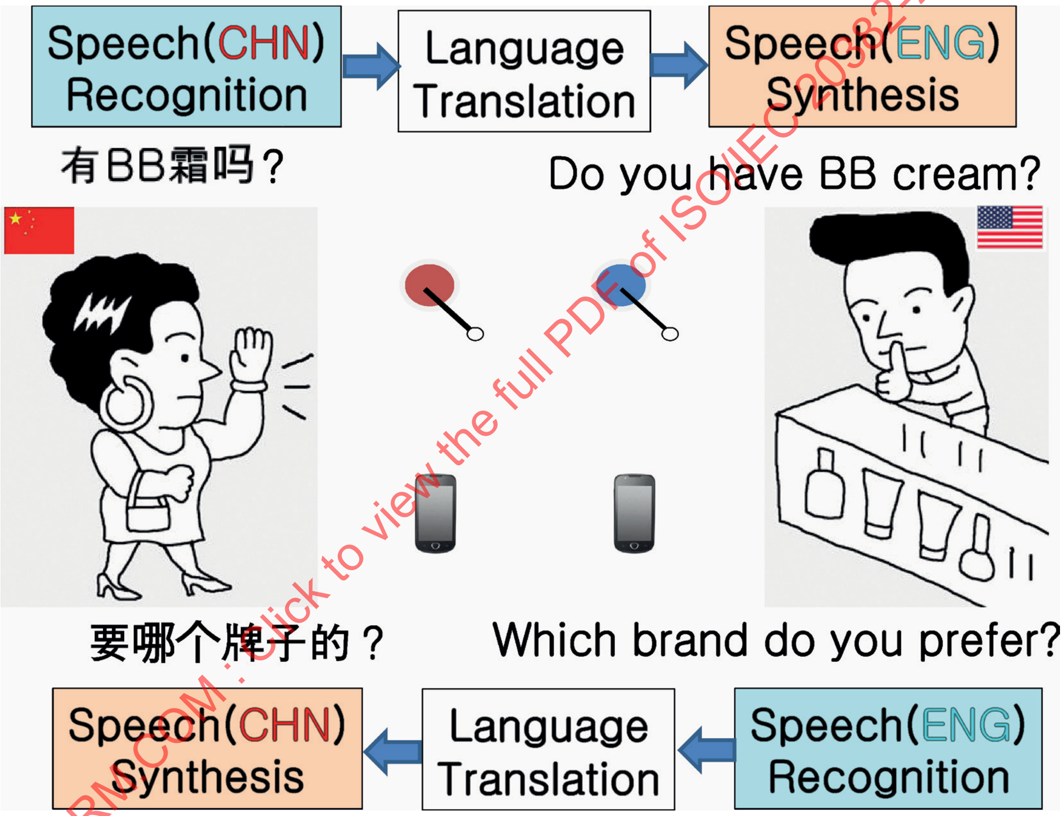


Figure A.1 — Functional components of F2F speech translation

A Chinese woman approaches an American cosmetics store and asks for BB cream in the Chinese language. The Chinese speech recognition module recognizes the Chinese speech and converts it into a Chinese sentence. The Chinese-to-English language translation module translates the Chinese sentence into English 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 recognizes the English speech and converts it into an English sentence. The English-to-Chinese language translation module translates the English sentence into a Chinese sentence with the same meaning. The Chinese speech synthesis module generates Chinese speech from the Chinese sentence.

In previous times, one hired a human translator with multi-lingual capability to accomplish this goal. Since the beginning of the 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 the components of speech translation. Various speech translation services, as well as application software for mobile devices, have appeared 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 document 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 the same language easily. The user interface of this document is just the same as that of communication among the people of the same language. This method can be called the “3rd generation” or zero-effort speech translation.

In Annex A, a comparison among the three generations of F2F speech translation is given in terms of system functions and performances.

A.2 1st generation: human translator

Until the 20th century, people hired a human translator to visit a place with a different language if they could not speak that language. The human translator in Figure A.2 tries his best to help the employer to do whatever the employer wants to do.

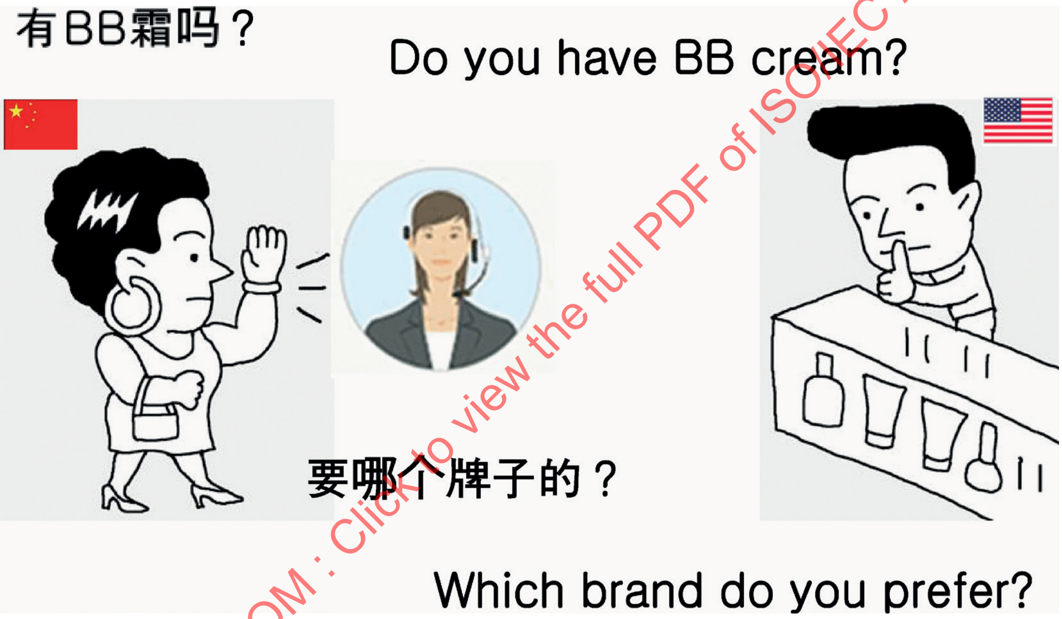


Figure A.2 — 1st generation F2F speech translation: human translator

It is very easy for the employer to translate his speech because he just needs to tell to the human translator what he wants to do. It takes time to communicate since the human translator needs to translate the sentence and talk 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 a 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 part of the conversation.

A.3 2nd generation: automatic speech translation

Since the beginning of the 21st century, automatic speech recognition, automatic language translation and speech synthesis have been developed, and the quality of these technologies is sufficient to serve as components of speech translation. Various speech translation services, as well as application software