International Standard

ISO/IEC 23773-3

First edition
2024-07

Information technology — User interfaces for automatic simultaneous interpretation systems —

Part 3: System architecture

Technologies de l'information — Interfaces utilisateur pour les systèmes d'interprétation simultanée automatique —

Partie 3: Architecture du système

Reference number
ISO/IEC 23773-3:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

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.

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 document 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 or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

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

For an explanation of 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 www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards..

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

A list of all parts in the ISO/IEC 23773 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Communication between users of different languages is a global trend that is increasing. Real-time, automatic simultaneous interpretation is needed for different applications such as video calls, live lecture translation and wearable translation devices. Market demands for real-time automatic simultaneous interpretation of free-style continuous utterances in the travel sector, global event management, as well as for phone calls, lectures or meetings, are also increasing. A standardized user interface (UI) for automatic simultaneous interpretation systems fulfils these different needs for communication.

ISO/IEC 23773-1 provides a general description of an automatic simultaneous interpretation system designed to interoperate among different natural languages for spontaneous speech and text.

While traditional speech-to-speech translation described in ISO/IEC 20382-1 and ISO/IEC 20382-2 addresses the functional equivalent of consecutive interpretation, this document focuses on the functional equivalent of simultaneous interpretation.

ISO/IEC 23773-2 provides the requirements and functional components for the UI of automatic simultaneous interpretation systems.

This document provides a reference architecture for automatic simultaneous interpretation systems including functional modules and communication interfaces in a high-level approach.

Information technology — User interfaces for automatic simultaneous interpretation systems —

Part 3: System architecture

1 Scope

This document provides a description of a system architecture for real-time automatic simultaneous interpretation systems for spontaneous speech designed to interoperate among different natural languages.

While traditional speech-to-speech translation addresses the functional equivalent of consecutive interpretation, this document focuses on the functional equivalent of simultaneous interpretation.

This document does not cover sign language interpretation.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

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

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

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

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

3.1
incremental knowledge learning
knowledge accumulated through learning from previous experiences in the context of translation

3.2
automatic simultaneous interpretation
automatic interpretation where the translation is performed continuously while the speaker speaks without waiting for the translation to finish sentence by sentence

Note 1 to entry: Input is speech or text, or both, and output is speech or text, or both.

Note 2 to entry: While interpretation deals with spoken language in real time, translation focuses on written content.

3.3
interpretation unit
unit of the user's utterance which is the target for interpretation in the simultaneous interpretation in order to make the continuous translation

4 Abbreviated terms

ALT alternative

APE automatic post-editing

DB database

DNN deep neural network

KB knowledge base

ML machine learning

NMT neural machine translation

POS part of speech

RBMT rule-based machine translation

RNN recurrent neural network

LAS listen and spell

LSTM long short-term memory

5 Architecture of simultaneous interpretation system

5.1 General

The automatic simultaneous interpretation system consists of the following functional components which are presented in Figure 1. Clause 5 describes, in detail, the functional components, their sub-modules and interfaces among them. Figure 2 presents the flow chart of the automatic simultaneous interpretation to show how the functional components interact in the interpretation process.

— Simultaneous interpretation application: The simultaneous interpretation application functional component is the top layer of the interpretation system that provides the interface between the interpretation service and the system components.

— Continuous speech recognition: Speech recognition is performed continuously on the speech units as sentence units or interpretation units from vocalized speech. The vocalized speech is input in real time posed by the speech recognition engine that uses an acoustic model and a language model. For more information of continuous speech recognition, see ISO/IEC 24661.

— Interpretation unit extraction: A real-time interpretation unit extraction functional component forms one or more of the speech units into an interpretation unit.

— Real-time simultaneous interpretation: The user speech is continuously translated into a target language based on the interpretation unit which is formed by the real-time interpretation unit extraction module to produce natural translation results without stopping.

— Incremental knowledge learning: Knowledge required for the speech recognition and translation is acquired from different knowledge sources such as speech data, user log data, user and domain data and on-line data. The knowledge is incrementally learned and structured into a speech/translation knowledge DB and a user/domain knowledge DB that will be accessed and used by the system processes.

— Presentation of translation results: The functional component of presentation of translation results processes the output of the interpretation system and manages the presentation control functions. The translation results are presented to the users in different output formats, such as text, speech, or gestures

including sign languages, for different devices depending on the service types that the interpretation system provides.

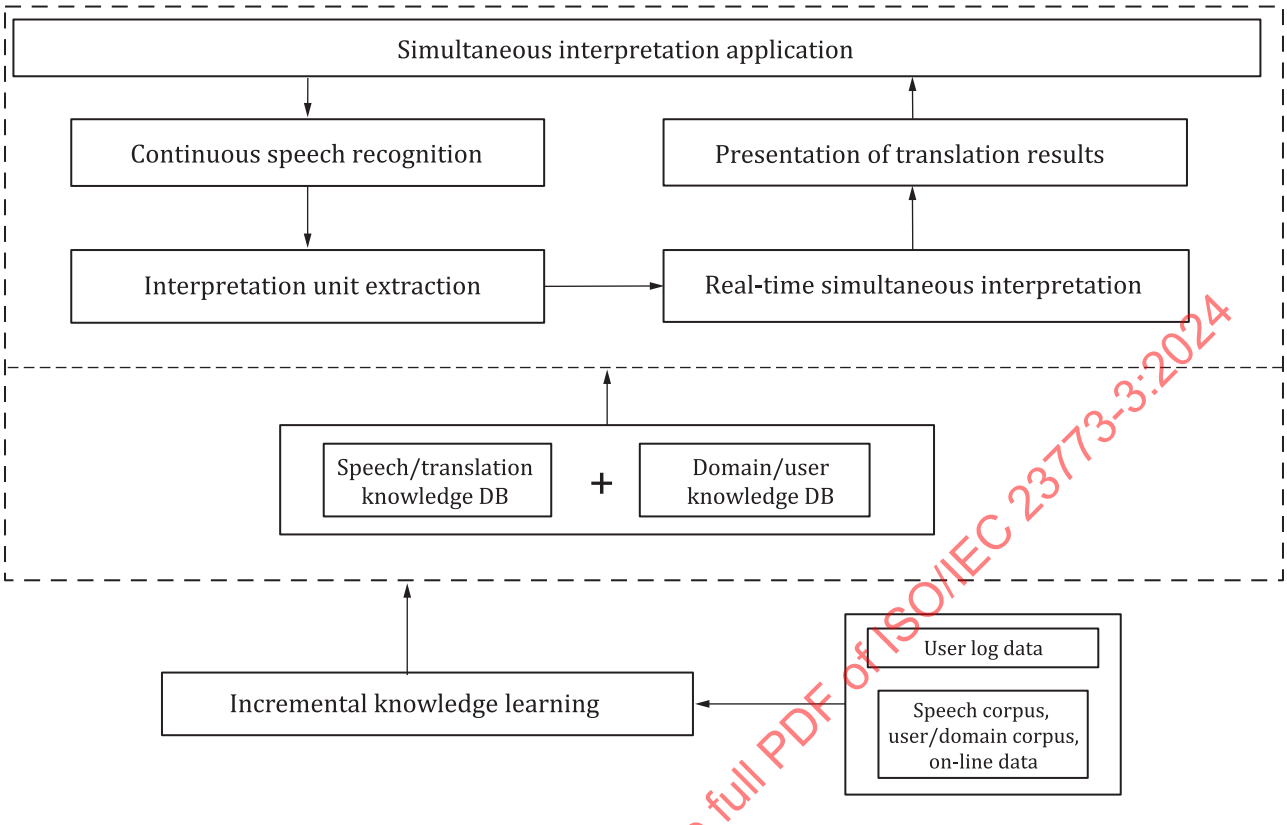

Figure 1 — Architecture of simultaneous interpretation system

ALT text for Figure 1

Figure 1 consists of three big boxes. The first box (Box 1) is the biggest one and it is located at the top of the figure and the other two boxes (Box 2 and Box 3) are located horizontally under the first box. The three big boxes are connected by arrows. One arrow goes from Box 2 (Incremental knowledge learning) to Box 1. Another arrow goes from Box 3 to Box 2.

Box 1 consists of several boxes. The top box is labelled as “Simultaneous interpretation application”. There are four boxes under the top box, two on the right side and two on the left side. The first box under the left side of the top box is labelled as “Continuous speech recognition”. The second box under the left side of the top box is labelled as “Interpretation unit extraction”. The two boxes under the “Simultaneous interpretation application” are “Presentation of translation results” and then the “Real-time simultaneous interpretation” box under it. The arrow goes from the top box to the first box on the left and an arrow goes from that box to the second box on the left. A horizontal arrow goes from the second box on the left to the “Real-time simultaneous interpretation” box and another arrow goes up to the first box on the right. Finally, an arrow goes from the first box on the right to the top box. There are two boxes at the bottom of Box 1. They are “Speech/translation knowledge DB” box and “Domain/user knowledge DB”. The two boxes are connected to each other with a plus “+” mark. There is also an arrow from the plus “+” mark to the upper boxes.

Under the big Box 1, there is Box 2 (Incremental knowledge learning) and it is connected by an arrow pointing to the big Box 1.

Finally, the last big box, Box 3 has two boxes vertically aligned: “User log data” box and “Speech corpus, user/domain corpus, on-line data” box. An arrow points from Box 3 to Box 2 (Incremental knowledge learning).

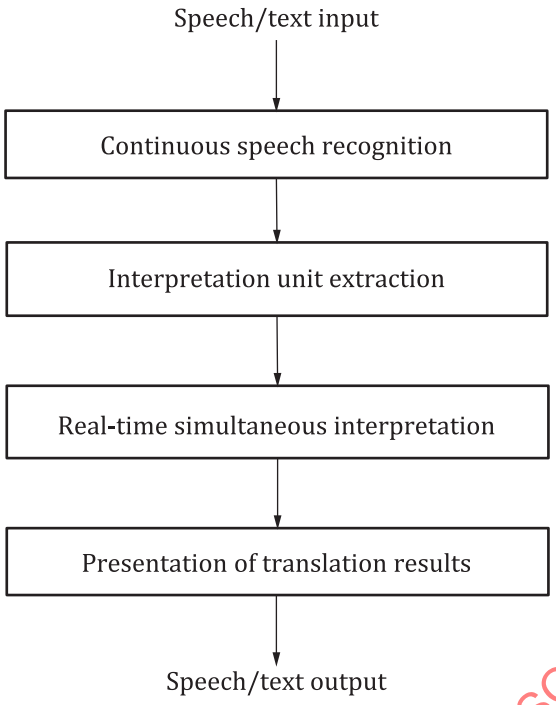

Figure 2 — Flowchart of simultaneous interpretation process

ALT text for Figure 2

Figure 2 consists of four horizontally-aligned boxes and text at the top and bottom. Arrows point downwards from the top text through the four boxes to the bottom text. The text on the top is “Speech/text input.” The first box under the top text is connected by a downward arrow and is labelled as “Continuous speech recognition". Likewise, the second box is labelled as “Interpretation unit extraction". The third box is labelled as “Real-time simultaneous interpretation”. The fourth box is labelled as “Presentation of translation results” which is connected to the bottom text “Speech/Text output”.

5.2 Continuous speech recognition module

5.2.1 General

Speech recognition is performed continuously on the speech input in real time posed by the speech recognition engine of the simultaneous interpretation system. The submodules of a continuous speech recognition module are a feature extraction/noise processing module, an utterance adaptation module and a multiple decoder.

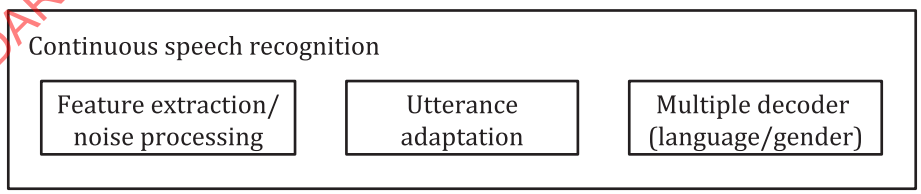

Figure 3 — Continuous speech recognition module

ALT text for Figure 3

Figure 3 consists of a big box which is labelled as “Continuous speech recognition”. It contains three horizontally-aligned small boxes which are titled: “Feature extraction/noise processing”, ”Utterance adaptation”, and “Multiple decoder (language/gender)”.

5.2.2 Feature extraction/noise processing

The feature extraction/noise processing sub-module extracts features from the input speech signal and removes noises. In addition, the feature extraction sub-module extracts voice characteristic information such as pitch, vocal intensity, speech speed and vocal tract characteristics of an original speech. The voice characteristic information is used in the personalized speech synthesis module to generate a synthetic sound with characteristics similar to those of an original speaker's voice.

5.2.3 Utterance adaptation

Utterance adaptation function aims to improve the speech recognition performance by applying a speaker's features in the deep learning acoustic model as a speaker embedding vector. Unsupervised utterance adaptation aims at improving the accuracy from an unknown speaker.

5.2.4 Multiple decoder

The multiple decoder function performs speech recognition and language/gender identification at the same time. The performance of the speech recognition is important because of the multiple function of the decoder. In the continuous speech recognition, the past information is constantly used as the context information. In this sense, the recurrent neural network (RNN)-based long short-term memory (LSTM) method is applied for the acoustic model training. For the end-to-end model, attention based DNNs such as listen and spell (LAS) and multi-head attention-based transformer model are applied.

5.2.5 Interfaces of submodules

Table 1 describes the input and output of each sub-module of the continuous speech recognition.

Table 1 — Input and output of sub-modules in continuous speech recognition

Function name	Input	Output
Feature extraction/ noise processing	Speech signal	Features, clean speech
Utterance adaptation	Acoustic model	Adapted acoustic model
Multiple decoder	Speech	Text

5.3 Interpretation unit extraction module

5.3.1 General

A real-time interpretation unit extraction functional module forms one or more of the speech units into an interpretation unit. It is necessary to extract optimal interpretation units from the input. The interpretation unit can be a phrase, a clause, or a sentence depending on the input. For example, a long speech consisting of several sentences can be spoken in one breath, a single sentence can be spoken with multiple breaths, speech could be left unfinished by a sentence, or meaningless interjections are frequently made. In these cases, a correct translation result cannot be generated using a conventional automatic translation methodology in which translation is performed for each sentence unit.

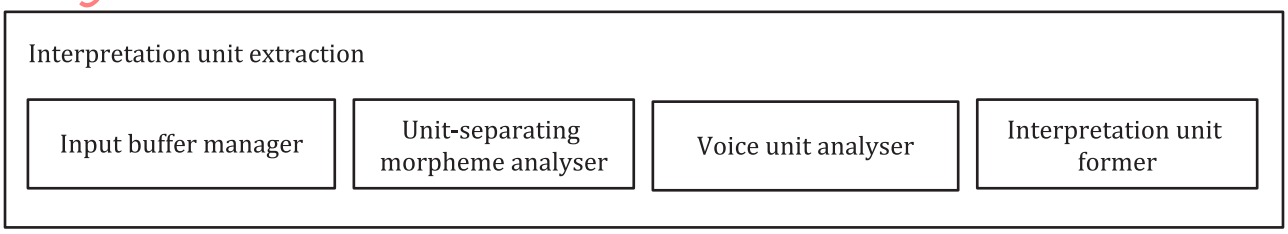

Figure 4 — Interpretation unit extraction module

ALT text for Figure 4

Figure 4 consists of a big box which is labelled as “Interpretation unit extraction” and contains four horizontally-aligned boxes which are titled: “Input buffer manager”, “Unit-separating morpheme analyser”, “Voice unit analyser” and “Interpretation unit former”.

5.3.2 Input buffer manager

The input buffer manager stores a voice unit that is input text from the speech recognition module. It also manages the remaining voice units that had not been included in the previous interpretation unit. The current voice unit and remaining voice unit will be processed together to form the interpretation unit.

5.3.3 Unit separating morpheme analyser

The unit separating morpheme analyser detects the morpheme of the input voice unit from the input buffer based on the POS analysis. The POS analysis results are, for example, noun, verb, pronoun.

5.3.4 Voice unit analyser

The voice unit analyser re-separates the voice unit based on the morpheme analysis results to validate the correct voice unit which will be included in the interpretation unit.

5.3.5 Interpretation unit former

The interpretation unit former combines the current voice unit and previous voice unit to form the interpretation unit. Voice units from the input buffer are combined or separated to form interpretation units, which are the units for correct translation. The information used for the interpretation unit former includes features such as lexical features, POS features, time features and acoustic features. Lexical features are words that show signals for the separation of an interpretation unit such as “For example”. The acoustic features include pauses, prosody and accents. These features are used independently or in combination of several features among them.

5.3.6 Interfaces of submodules

Table 2 describes the input and output of each sub-module of the interpretation unit extraction module.

Table 2 — Input and output of sub-modules in interpretation unit extraction

Function name	Input	Output
Input buffer manager	Text	Voice unit
Unit separating morpheme analyser	Voice unit	Voice unit with morpheme annotation
Voice unit analyser	Voice unit with morpheme annotation	New voice unit
Interpretation unit former	New voice unit	Interpretation unit

5.4 Real-time simultaneous interpretation module

5.4.1 General

The user speech is continuously translated into a target language based on the interpretation unit which is formed by the real-time interpretation unit extraction module to produce natural translation results without stopping. The translation is performed by rule-based translation using language analysis, context-based translation, and translation based on machine learning (ML) methods such as DNN. Those different approaches are combined as the hybrid translation and final translation results are produced. The sub-modules are described in Figure 5 and the flow chart of real-time simultaneous interpretation is presented in Figure 6.

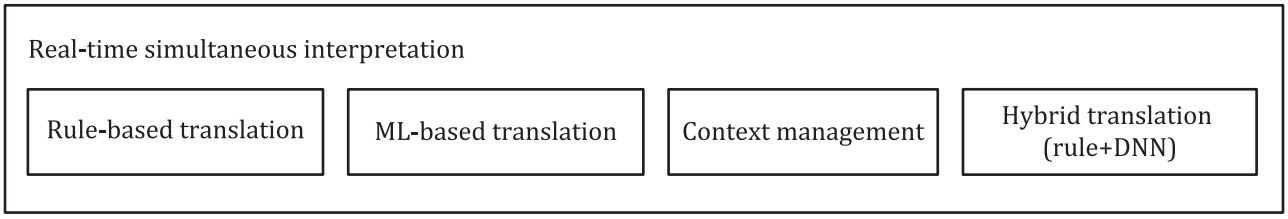

Figure 5 — Real-time simultaneous interpretation module

ALT text for Figure 5

Figure 5 consists of a big box which is labelled as "Real-time simultaneous interpretation" and contains four horizontally-aligned boxes which are titled: "Rule-based translation", "ML-based translation", "Context management" and "Hybrid translation (rule+DNN)".

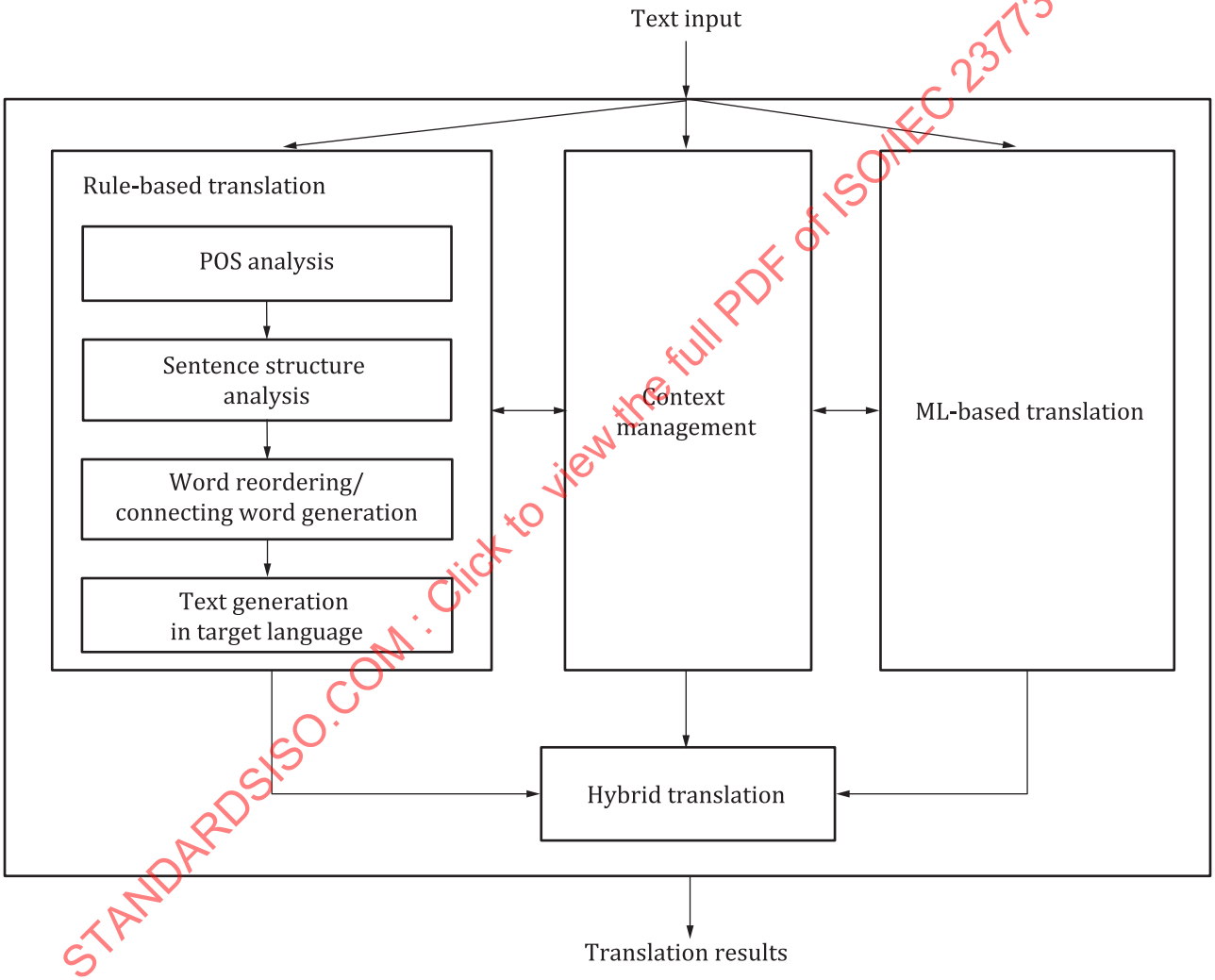

Figure 6 — Flowchart of real-time simultaneous interpretation

ALT text for Figure 6

Figure 6 consists of one big box with text "Text input" on the top and text "Translation results" at the bottom. The two texts are connected with downward vertical arrows with a big box in the middle. The big box has four different sub boxes among which three are on the upper part [i.e. Boxes 1 (Rule-based translation), 2 (Context Management), 3 (ML-based translation)] and one smaller Box 4 (Hybrid translation) at the lower part. From the middle point of the top in the big box, there are three arrows pointing downwards to connect to the three sub boxes. The first sub box is labelled as "Rule-based translation" and has four vertically aligned

boxes which are connected with downwards arrows: they are labelled as, “POS analysis”, “Sentence structure analysis”, “Word reordering/connecting word generation” and “Text generation in target language”. The four boxes are connected to another box labelled as “Context management” with four bidirectional arrows. The “Context management” box is connected to the “ML-based translation” box on the right with a bi-directional arrow. From Box 1 (Rule-based translation), Box 2 (Context Management) and Box 3 (ML-based translation), there are three downward arrows connecting to Box 4 (Hybrid translation).

5.4.2 Rule-based translation

The text from the interpretation unit extraction module is sent to rule-based translation, ML-based translation and the context management module as input. In the rule-based translation module, the input text is processed through language analysis steps such as POS analysis, sentence structure analysis, and word reordering and connecting word generation, then finally text generation in target language. The translation results are fed into the hybrid translation module to combine with the ML-based translation results.

5.4.3 Machine learning (ML)-based translation

The text input from the interpretation unit extraction module is processed through ML translation and the intermediate translation results are generated. The intermediate translation results are fed into the hybrid translation module to combine with the rule-based translation results.

As the various DNN algorithms develop, more end-to-end translation systems are emerging. In cases that the end-to-end neural machine translation (NMT) independently applies, input text in the source language is processed with a memory to store a real time interpretation and translation program for the input token. The processor to execute the translation program combines an output of a translation network with an output of an action network to compose a final translation result in the communication module of the interpretation application.

5.4.4 Context management

The context management sub-module stores the previous input sentences with language analysis results and translation results. That information is provided for the translation module as context for correct analysis and for enhanced translation by error correction to reflect in the subsequent interpretation units.

5.4.5 Hybrid translation (rule+DNN)

In the hybrid translation sub-module, both rule-based translation results and ML (such as DNN) translation results are combined to generate the final translation results. Hybrid machine translation technology combines an NMT system with a well-developed rule-based machine translation (RBMT) system to utilize the stability of RBMT to compensate for the inadequacy of NMT in rare resource domains. A series of experiments showed that the translation accuracy of the hybrid system was improved especially in out-of-domain translations where training data are not enough for NMT.

5.4.6 Interfaces of submodules

Table 3 describes the input and output of each sub-module of the real-time simultaneous interpretation module.

Table 3 — Input and output of sub-modules in real-time simultaneous interpretation

Function name	Input	Output
Rule-based translation	Text	Text in target language
ML-based translation	Text	Text in target language
Context management	Previous results	Context information
Hybrid translation	Tagged text in source language	Text in target language

5.5 Incremental knowledge learning module

5.5.1 General

Knowledge required for the speech recognition and translation is acquired from different knowledge sources such as speech data, user log data, user and domain data and on-line data. The knowledge is incrementally learned and structured into a speech/translation knowledge DB and a user/domain knowledge DB that will be accessed and used by the system processes.

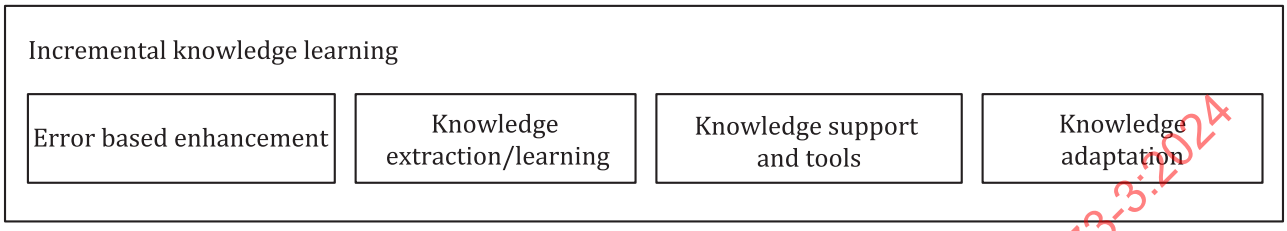

Figure 7 — Incremental knowledge learning module

ALT text for Figure 7

Figure 7 consists of a big box which is labelled as “Incremental knowledge learning” and contains four horizontally-aligned boxes which are titled: “Error-based enhancement”, “Knowledge extraction/learning”, “Knowledge support and tools” and “Knowledge adaptation”.

5.5.2 Error-based enhancement

The error-based enhancement module finds errors included in the translation results of the automatic interpretation module. It corrects the errors and generates edited sentences. This function is usually implemented as automatic post-editing (APE) for translation output as the post processing module. The APE process receives the source sentence and translated sentence at the same time as the input and constructs a sequence-to-sequence model to reflect in the translation DNN model. The error correction data by the user after the speech recognition or translation is also reflected in the knowledge base (KB) by the error-based enhancement module.

5.5.3 Knowledge extraction/learning

It is critical to collect spontaneous speech data for continuous speech recognition. Radio interviews and broadcasting text are examples of good candidates for spontaneous speech feature extraction. For simultaneous interpretation, diverse domain data are needed to minimize the out of vocabulary (OOV) such as IT, computer, social media, politics, education and art. Those data are classified into sub areas and built into text DBs. They are also used for the data sets to build a language model to learn N-Gram, which is a sequence of adjacent words, for different domains.

5.5.4 Knowledge support and tools

For simultaneous interpretation, the interpretation unit-based corpus is most highly in demand. The interpretation unit-based corpus is composed of pairs of sentences in two different languages where the interpretation units are recognized and translated sentences are aligned with the source sentences by the interpretation unit. To construct this aligned corpus, word alignment, sentence segmentation, and reordering of the interpretation units should be performed.

The newly added KB data helps to improve translation accuracy. Data for the KB can also be collected by the following tools:

— News search and crawling to input specific keywords to initiate a targeted search for relevant news articles. Subsequently, the system engages in crawling these articles through their respective URLs, systematically extracting valuable information.

— Sentence segmentation and translation to segment the collected data by sentence, to correct the translation results of the sentence and to store in the KB as a parallel corpus to be used for the DNN training data.

— Terminology extraction to automatically extract terminology according to the frequencies from the sentence-based news contents. If the translated terms exist in the document, the pair of words and translated words are stored. Otherwise the translated terms should be added by searching in the on-line dictionaries. The generated list of translation words with their original terms will be used in the translation module b learning mechanism.

— Text extraction from documents to extract text from documents and do the segmentation by the sentence unit. The next step is to extract new words and proper nouns and construct dictionary to be used for the translation module.

5.5.5 Knowledge adaptation

Knowledge adaptation function aims to improve the performance when the new domain or user data are not enough for the translation function. The existing data for other domains and different users are utilized after being modified accordingly. Transfer learning is a method to adapt to a new domain with sparse training data using the existing data resources from other domains.

5.5.6 Interfaces of submodules

Table 4 describes the input and output of each sub-module of the incremental knowledge learning module.

Table 4 — Input and output of incremental knowledge learning

Function name	Input	Output
Error based enhancement	KB	Enhanced KB
Knowledge extraction/learning	Data	KB/language model
Knowledge support and tools	KB/tools	KB/tools
Knowledge adaptation	Knowledge data	Adapted knowledge data

5.6 Presentation of translation results

5.6.1 General

The functional component of the presentation of translation results processes the output of the interpretation system and manages the presentation control functions. The translation results are presented to the users in different output formats, such as text, speech, or gestures, including sign languages, for different devices depending on the service types that the interpretation system provides.

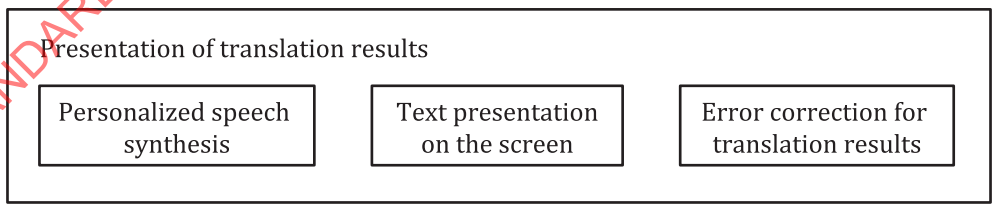

Figure 8 — Presentation of translation results

ALT text for Figure 8

Figure 8 consists of a big box which is labelled as “Presentation of translation results” and contains three horizontally-aligned boxes which are titled: “Personalized speech synthesis”, “Text presentation on the screen”, and “Error correction for translation results”.