INTERNATIONAL STANDARD

ISO/IEC 24661

First edition
2023-05

Information technology — User interfaces — Full duplex speech interaction

Technologies de l'information — Interfaces utilisateur — Interaction vocale en duplex intégral

Reference number
ISO/IEC 24661:2023(E)

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).

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) or the IEC list of patent declarations received (see https://patents.iec.ch).

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*.

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

Speech interaction user interface (UI) has been widely used for industrial applications and daily services. For example, it can be applied to automatic customer service in the telecommunication industry as a part of an interactive voice response system. From a communication point of view, a speech interaction UI can be recognized as a duplex-based system which enables bidirectional communication. In the early stages, speech interaction UIs for conventional dialogue systems were generally half duplex (HDX) based and were designed to be in a turn-oriented work mode. As the requirements of human-machine interaction have grown in complexity and diversity, the turn-oriented speech interaction UI has become unfit for a conversation between humans and machines.

Currently, full duplex (FDX) techniques are used in the speech interaction UI to support session-oriented conversations between humans and machines. The most significant differences between turn-oriented and session-oriented speech interactions are continuity and naturalness, which have made great progress in various applications of speech interaction UI, e.g. smart speaker, chatbot, intelligent assistant.

In recent years, a growing number of FDX speech interaction UIs have been studied and developed. This requires a common understanding of general models and specifications through standardization activities. In response to the standardization needs both from industry and academia, this document intends to provide a reference architecture, functional components and technical requirements of FDX speech interaction UI. For the benefit of system designers, developers, service providers and ultimate users, this document is composed of the following clauses:

— Clause 5 describes a functional view and general features of FDX speech interaction;

— Clause 6 provides a reference architecture and functional layers of FDX speech interaction UI;

— Clause 7 specifies the functional requirements regarding each functional layer;

— Clause 8 discusses the processes of FDX speech interaction UI;

— Clause 9 describes security and privacy considerations related to FDX speech interaction UI.

Information technology — User interfaces — Full duplex speech interaction

1 Scope

This document specifies user interfaces (UIs) designed for full duplex (FDX) speech interaction. It also specifies the FDX speech interaction model, features, functional components and requirements, thus providing a framework to support natural conversational interfaces between humans and machines. It also provides privacy considerations for applying FDX speech interaction.

This document is applicable to UIs for speech interaction and communication protocols for setting up a session-oriented FDX interaction between humans and machines.

This document does not define the speech interaction engines themselves or specify the details of specific engines, devices and approaches.

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
duplex
method of communication capable of transmitting data in both directions

[SOURCE: ISO 21007-1:2005, 2.18]

3.2
full duplex
FDX
method of communication capable of transmitting data in both directions at the same time

[SOURCE: ISO 21007-1:2005, 2.25]

3.3
functional unit
entity of hardware or software, or both, capable of accomplishing a specified purpose

Note 1 to entry: Functional units can be integrated as a system.

[SOURCE: ISO/IEC 2382:2015, 2123022, modified — Note 1 to entry has been changed and Note 2 and 3 to entry have been removed.]

3.4
half duplex
HDX
method of communication capable of transmitting data in both directions but only in one direction at any time

[SOURCE: ISO 21007-1:2005, 2.27]

3.5
microphone array
system that is composed of multiple microphones with definite spatial topology, which samples and filters the spatial characteristics of signals

3.6
speech interaction
activities of information transmission and communication between humans and a system through speech

Note 1 to entry: A system can be seen as a combination of *functional units* (3.3).

3.7
speech recognition
automatic speech recognition
ASR
conversion, by a *functional unit* (3.3), of a speech signal to a representation of the content of the speech

Note 1 to entry: The content to be recognized can be expressed as a proper sequence of words or phonemes.

[SOURCE: ISO/IEC 2382:2015, 2120735, modified — Notes 2 to 4 to entry have been removed.]

3.8
speech synthesis
generation of speech from data through a mechanical method or electronic method

Note 1 to entry: Speech can be generated from text, image, video and audio. The process of conversion from text to speech is the main approach in *speech interaction* (3.6).

Note 2 to entry: The result of speech synthesis is also called "artificial speech" in order to differ from natural speech through human vocal organs.

3.9
voice activity detection
VAD
process of analysis and identification of the starting and ending points of valid speech in a continuous speech stream

3.10
voice trigger
process in a system in the audio stream monitoring state, which switches to command word recognition, continuous speech recognition and other processing states after the detection of certain features or events

4 Symbols and abbreviated terms

AAC	advanced audio coding
AC3	audio coding 3
AI	artificial intelligence
ASR	automatic speech recognition
EVRC	enhanced variable rate codec
FDX	full duplex
HDX	half duplex
ML	machine learning
MP3	MPEG audio layer 3
NER	named entity recognition
NLG	natural language generation
NLP	natural language processing
NLU	natural language understanding
SNR	signal-to-noise ratio
TTS	text-to-speech
UI	user interface
VAD	voice activity detection
WAV	waveform audio file format
WMA	Windows media audio

5 Overview of FDX speech interaction UI

5.1 Functional view

Speech interaction UI can function as a communication channel between a human and a system. A user can apply a speech interaction UI to have a conversation with a system, while a system can also respond to the user with synthesized speech through the speech interaction UI. Such bidirectional communication can be viewed as a duplex speech interaction. With different data transmission sequences, there are two types of duplex speech interactions, including HDX mode and FDX mode.

In the case of HDX speech interaction, both a human and a system can communicate with each other in one direction at a time. An HDX speech interaction is characterized as a turn-oriented dialogue, where a system will return to the default state after it finishes one round of dialogue. In addition, the system cannot collect speech signals during the process of its speech broadcasting.

NOTE 1 A typical HDX-based communication system is a two-way radio such as walkie-talkie. A walkie-talkie uses a "push-to-talk" button to control the signal transmission channel. A user can turn on the transmitter and turn off the receiver by using the button, so that the voice from remote users cannot be heard.

In contrast to HDX speech interaction, FDX speech interaction allows a human and a system to communicate with each other simultaneously. An FDX speech interaction is characterized as a session-oriented conversation, where a system keeps the conversation continuous and ensures that both user and system are in the same context after two or more rounds of dialogue. In addition, the user and the system can speak within the same interval of time. Example scenarios of FDX speech interaction are shown in Annex A.

NOTE 2 A typical FDX-based communication system is the telephone, where both local and remote users can speak and be heard at the same time.

From the functional point of view, a system can keep receiving the input data from the user and providing feedback to them through an FDX speech interaction UI during the whole human-machine conversation. Figure 1 depicts a functional view of FDX speech interaction UI that includes inputs, processing and outputs.

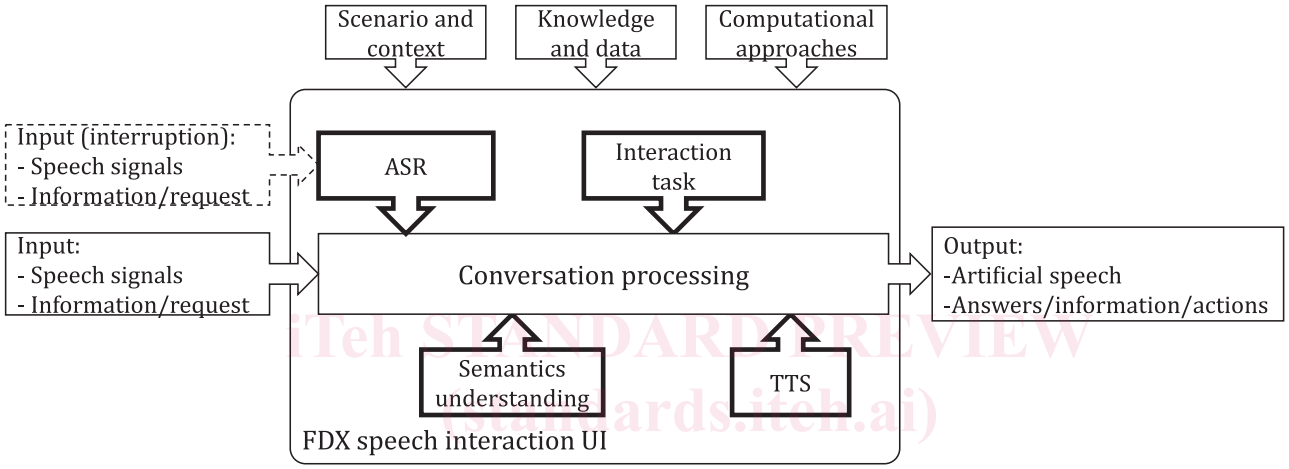

Figure 1 — Functional view of FDX speech interaction UI

This functional view provides a non-technical description of how an FDX speech interaction uses UI to achieve its goal. Through the FDX speech interaction UI, a system can receive the input speech signals, transcript the useful signals into the text, abstract the semantic information from transcription text, make predictions and decisions regarding interaction tasks based on semantic information, and either take actions based on the decisions or provide speech feedback to users as the outputs, or both. In contrast to HDX mode, an FDX speech interaction is characterized by functions of continuous speech acquisition by a system after it has been awoken once. Such function can be performed even when a system is outputting synthesized speeches or other actions. This is considered to be a conversational interruption, i.e. technically, both uplink speech stream and downlink speech stream may take place at the same time. An FDX speech interaction UI shall have abilities to execute the conversation processing whenever there are speech interruptions and to generate updated outputs based on the new inputs.

During this process, scenarios and contexts can be used to define the semantic range of the conversation. A conversation can be cross scenarios and contexts. General knowledge and big data are required for the conversation processing. Computational approaches such as cloud computing and AI computational approaches should be introduced in the FDX speech interaction UI. Such functional components are applied to performing intelligent conversation processing, which is a distinguishing characteristic of FDX mode compared with HDX mode

5.2 Main characteristics

5.2.1 General

To demonstrate the breadth of FDX speech interaction UI, some common characteristics are described in 5.2.2 to 5.2.8. In the aggregate, these characteristics are intrinsic to many FDX speech interaction UIs, which will differentiate FDX speech interaction UIs from non-FDX speech interaction UIs. The list

of characteristics of FDX speech interaction UIs is not exhaustive, but broadly conceptual and not tied to a specific methodology or architecture.

5.2.2 Continuous

Through an FDX speech interaction UI, a user can keep talking as continuous inputs, while the system can keep receiving and processing the input data.

5.2.3 Natural

An FDX speech interaction UI can support a natural conversation between a human and a system. A system only needs to be awoken once at the beginning of the conversation. A user can talk at will and freely interrupt the system at any time during a conversation.

5.2.4 Adaptable

An FDX speech interaction UI can adapt to different changes in itself and the environment in which it is deployed. It can be used in different vertical industries and applied to cross-domain applications and tasks by feeding on dynamic data and updating status based on new data.

5.2.5 Initiative

An FDX speech interaction UI can exhibit dynamic predictions of conversational intention based on external data sources, control the pace of conversation, and actively provide feedback to guide the user for further steps.

5.2.6 Context-based

An FDX speech interaction UI builds its core functions on context, e.g. semantic understanding, historical information inheritance, data analysis and dialogue generation.

5.2.7 Knowledge-based

An FDX speech interaction UI can use knowledge from multiple sourced information, including contextual information, historical information, retrieval information and user information. This information can be stored in the general knowledge and database.

NOTE Retrieval information refers to information that is searched from other resources, e.g. internet website, database and knowledge base.

5.2.8 Model-based

An FDX speech interaction UI operates with various degrees of utilization of an acoustic model and language model. With the rapid development of emerging technologies, some FDX speech interaction UI are also embedded with cloud frameworks and AI-related models, e.g. convolutional neural network (CNN), recurrent neural network (RNN) and long short-term memory (LSTM) network.

6 Reference architecture of FDX speech interaction UI

6.1 General

Based on the functional view described in Figure 1, a reference architecture of FDX speech interaction UI is represented in terms of functional layers depicted in Figure 2. It provides a common understanding of function units and their relationships, which are technically necessary to construct an FDX speech interaction UI. While this reference architecture is not limited to a specific base technology (e.g. FDX speech interaction UI built with neural networks), it does not encompass every type of dynamic FDX speech interaction UI.

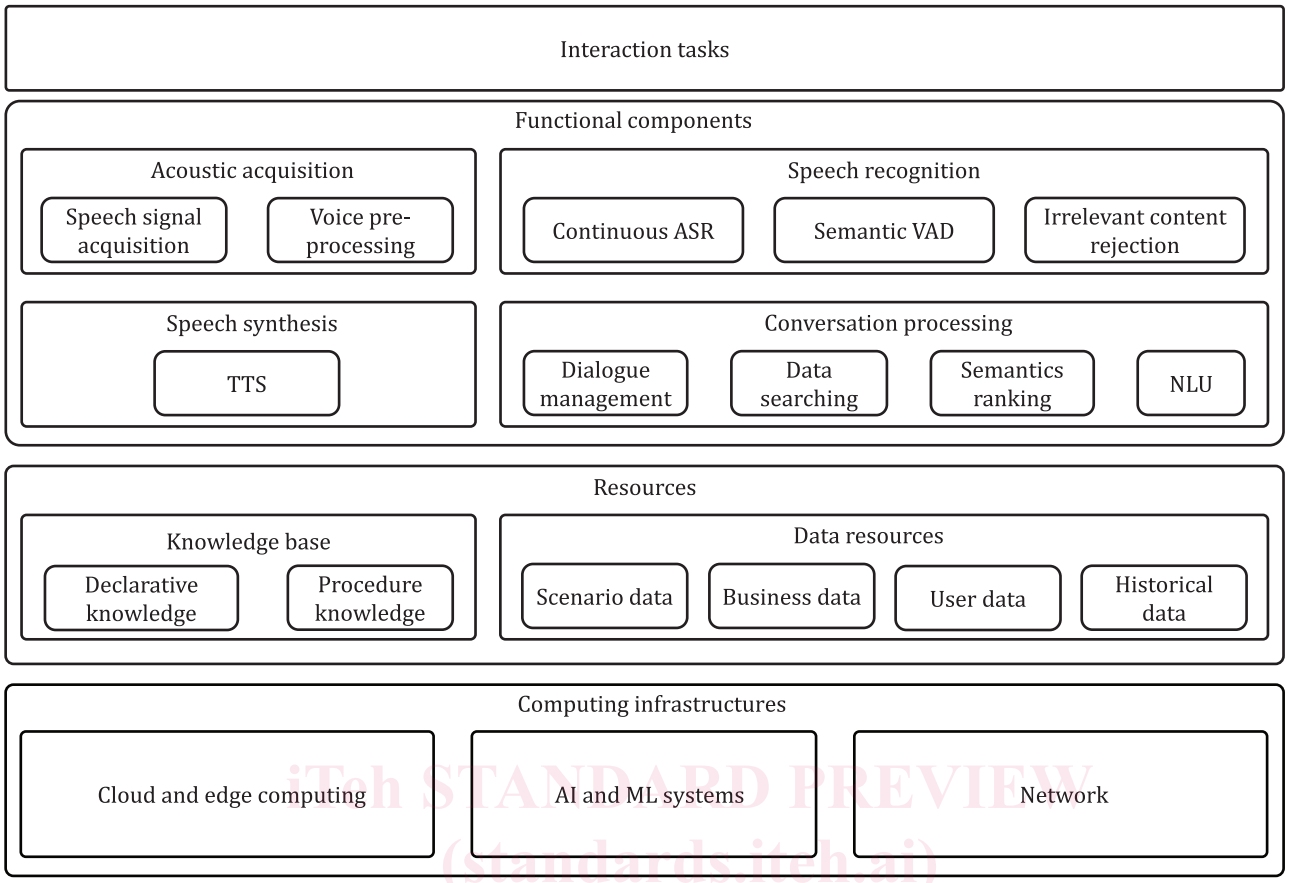

Figure 2 — Reference architecture of FDX speech interaction UI

This reference architecture consists of multiple layers and components. Such layers can be described in terms of the inputs, the outputs and the intents or functions. Each layer and its components can be used and tested separately. All layers can be integrated together to enable users to have conversations with the system and help to fulfil their requirements.

NOTE The system can be various smart devices, e.g. smart phone, smart home appliance, intelligent assistant app and customer service robot.

Speech data streams are transmitted through two physical channels. The upstream channel transmits speech data from the user to system. The downstream channel transmits speech data from the system to the user. Both channels shall be able to work at the same time without mutual-interference, thus to provide the system with the capability of "hearing" while "talking".

6.2 Interaction tasks

Interaction tasks refer to some specific purposes and requirements that need to be satisfied using an FDX speech interaction UI. One or more tasks can be defined for FDX speech interaction UI.

Each interaction task can be logically designed using traditional software engineering approaches, which involves defining the scenarios, the environmental features, the input and output, the function units, the database and the data flow.

Interaction tasks differ in the types of scenarios and the user requirements. Examples of interaction tasks can include, e.g. phone call, navigation, home service, chatting. While the scenarios should be defined in a general design, methods to resolve the specific problems should be addressed during the construction process. For example, using FDX speech interaction for a task of navigation, while a car driving scenario should be defined in the top-level design process, the point of interest (POI),