DRAFT INTERNATIONAL STANDARD

ISO/IEC DIS 24661

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

Voting begins on:
2022-04-20

Voting terminates on:
2022-07-13

Information technology — User interfaces — Full duplex speech interaction

ICS: 35.240.20

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.

Reference number
ISO/IEC DIS 24661:2022(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 iv

Introduction v

1 Scope 1

2 Normative references 1

3 Terms and definitions 1

4 Symbols and abbreviated terms 2

5 Overview of FDX speech interaction UI 3

5.1 Functional view 3

5.2 Main characteristics 4

5.2.1 General 4

5.2.2 Continuous 4

5.2.3 Natural 5

5.2.4 Adaptable 5

5.2.5 Initiative 5

5.2.6 Context-based 5

5.2.7 Knowledge-based 5

5.2.8 Model-based 5

6 Reference architecture of FDX speech interaction UI 5

6.1 General 5

6.2 Interaction tasks 6

6.3 Functional components 7

6.3.1 General 7

6.3.2 Acoustic acquisition 7

6.3.3 Speech recognition 9

6.3.4 Conversation processing 10

6.3.5 Speech synthesis 12

6.4 Resources 12

6.4.1 Knowledge base 12

6.4.2 Data resources 13

6.5 Computing infrastructures 13

6.5.1 Cloud and edge computing 13

6.5.2 AI and ML systems 14

6.5.3 Network 14

7 Functional requirements of FDX speech interaction UI 14

7.1 General requirements 14

7.2 Interaction tasks requirements 14

7.3 Functional components requirements 15

7.3.1 Acoustic acquisition requirements 15

7.3.2 Speech recognition requirements 15

7.3.3 Conversation processing requirements 16

7.3.4 Speech synthesis requirements 16

7.4 Resources requirements 17

7.5 Computing infrastructures requirements 17

8 Processes of FDX speech interaction UI 17

8.1 General 17

8.2 Engineering process 17

8.3 Interaction process 18

9 Security and privacy considerations of FDX speech interaction UI 20

Annex A (informative) **Example scenarios of FDX speech interaction** 21

Bibliography 23

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

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 ISO 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 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 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.

This document was prepared by 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.

Introduction

Speech interaction user interface 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 interactive voice response system. From the view of communication, a speech interaction user interface can be recognized as a duplex based system which enables bidirectional communication. In early stage, speech interaction user interfaces for conventional dialogue system were generally half duplex based and were designed to be turn–oriented work mode. As the growing complexity and diversity of requirements of human-machine interaction, the turn–oriented speech interaction user interface has become unfit for a conversation between human and machine.

Nowadays, full duplex (FDX) techniques are used in the speech interaction user interface to support session–oriented conversation between human and machine. The most differences between turn-oriented and session-oriented speech interaction are continuity and naturalness, which make great progress in various applications of speech interaction user interface, such as smart speaker, chatbot, intelligent assistant, etc.

In recent years, a growing number of FDX speech interaction user interfaces have been studied and developed. It requires a common understanding of general model 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 user interface. For the benefit of system designers, developers, service providers and ultimate users, this document is composed of the following clauses:

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

— Clause 6 provides a reference architecture and functional layers of FDX speech interaction user interface,

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

— Clause 8 discusses the processes of FDX speech interaction user interface,

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

This document is not intended to specify the very details of specific engines, devices and approaches.

Information technology — User interfaces — Full duplex speech interaction

1 Scope

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

This document is applicable to user interfaces for speech interaction and communication protocols for setting up a session–oriented full duplex interactions between human and machine. This document is not applicable to defining the speech interaction engines themselves.

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 terminological 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
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
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 human and system through speech

Note 1 to entry: A system can be seen as combination of functional units.

3.7
speech recognition
automatic speech recognition
conversion, by a functional unit, 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 - Note 2 to 4 to entry have been removed and admitted terms have been added.]

3.8
speech synthesis
generation of speech from data through 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.

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

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

3.10
voice trigger
speech wake-up
process that system, in the state of audio stream monitoring, switches to command word recognition, continuous speech recognition and other processing state 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 human and system. A user can apply a speech interaction UI to have a conversation with system, while 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 sequence, there are two types of duplex speech interaction including HDX mode and FDX mode.

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

NOTE 1 A typical HDX based communication system is the 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 human and system to communicate with each other simultaneously. An FDX speech interaction is characterized as a session-oriented conversation, where 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.

NOTE 2 A typical FDX based communication system is the telephone, 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 feedbacks 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.

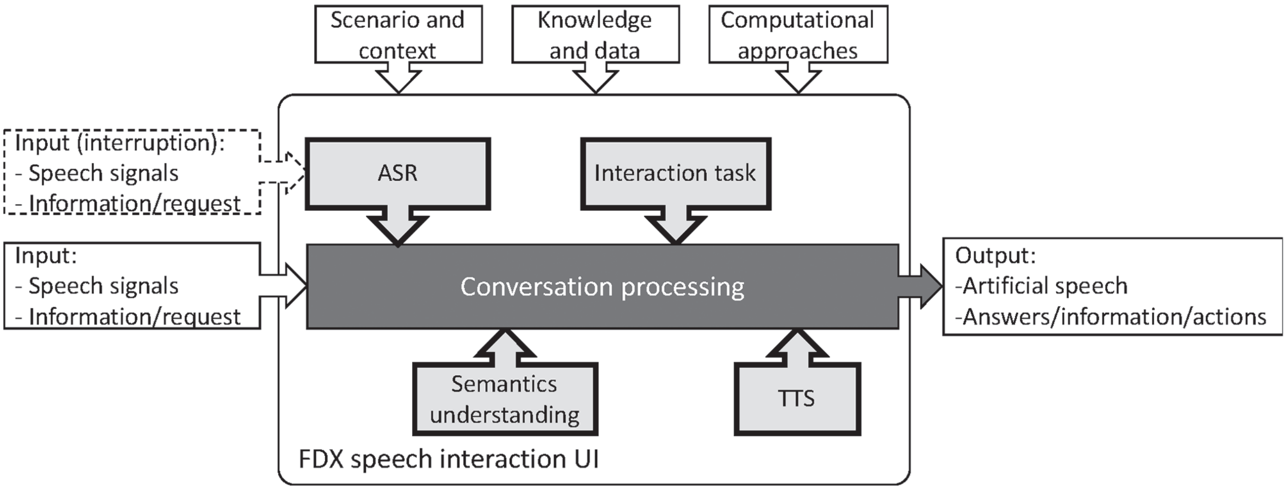

Figure 1 — Functional view of FDX speech interaction UI

This functional view provides a non-technical description of how an FDX speech interaction 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 to interaction tasks based on semantic information, take actions based on the decisions and/or provide speech feedbacks to users as the outputs. Different from HDX mode, an FDX speech interaction is characterized with functions of continuous speech acquisition by a system after it has been awoken once. Such function can be performed even when system is outputting synthesized speeches or other actions. This is behaved as a conversational interruption, which means, 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 defined the sematic 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 distinguished 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 as follows. 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 system can keep receiving and processing the input data.

5.2.3 Natural

An FDX speech interaction UI can support a natural conversation between human and system. System only needs to be awoken once at beginning of the conversation. A user can talk at will and freely interrupt 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 prediction of conversational intention based on external data sources, control the pace of conversation, and actively provide feedbacks to guide the user for the further steps.

5.2.6 Context-based

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

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 data base.

NOTE Retrieval information refers to those are searched from other resources such as internet website, database and knowledge base etc.

5.2.8 Model-based

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

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 the relationships of themselves, which are technically necessary to construct an FDX speech interaction UI. While this reference architecture is not limited to a specific base technology (such as FDX speech interaction UI built with neural networks), it does not encompass every type of dynamic FDX speech interaction UI.