

ICS 35.160
CCS L 62



**NATIONAL STANDARD OF THE
PEOPLE'S REPUBLIC OF CHINA**

GB/Z 177.8-2026

**Intelligence grading of artificial intelligence terminal - Part 8:
Speaker**

人工智能终端智能化分级 第8部分：音箱

Issued on: April 30, 2026

Implemented on: October 1, 2026

**Issued by: State Administration for Market Regulation;
Standardization Administration of China**

Table of Contents

Foreword	III
Introduction	IV
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Abbreviated terms	2
5 Key capabilities	2
5.1 Overview	2
5.2 L1 response level	2
5.3 L2 tool level	3
5.4 L3 assistance level	5
6 Determination of the level	6
Annex A (normative) Test methods	7
A.1 Test environment	7
A.2 L1 response level	7
A.3 L2 tool level	9
A.4 L3 assistance level	11
Annex B (informative) Typical use scenarios	16
Annex C (informative) Method of designing test scenarios	17
Bibliography	18

1 Scope

This document specifies the division into levels of the key intelligence capabilities of the smart speaker and the determination of the level, and gives the test methods.

This document applies to guiding the intelligence upgrading of smart speakers, and provides a reference for the design, development, application, selection and testing of artificial intelligence speakers.

2 Normative references

The following documents contain provisions which, through normative reference in this text, constitute indispensable provisions of this document. For dated references, only the edition cited applies. For undated references, the latest edition, including any amendments,

applies.

GB/Z 177.1-2026, Grading of intelligence for artificial intelligence terminals - Part 1:
Reference framework

GB/Z 177.2-2026, Grading of intelligence for artificial intelligence terminals - Part 2:
General requirements

3 Terms and definitions

The terms and definitions given in GB/Z 177.1-2026 and GB/Z 177.2-2026 and the following apply to this document.

3.1 speech wakeup, voice trigger - the process by which a speech interaction system in the state of monitoring an audio stream, on detecting the appearance of a particular feature or event, switches to another processing state such as command-word recognition or continuous speech recognition. [Source: GB/T 36464.2-2018, 3.13]

3.2 wake word - the particular word or phrase by which the user wakes the device and starts the speech interaction.

3.3 speech recognition - the process of converting the sound signal of a human being into text or into an instruction. [Source: GB/T 21023-2007, 3.1]

3.4 speech synthesis - the process of synthesizing human language by mechanical or electronic means. Note: the speech produced by this process is called synthetic speech, as distinct from the natural speech produced by the human vocal organs, and is sometimes also called artificial speech. [Source: GB/T 21024-2007, 3.1]

5 Key capabilities

5.2.2.3 Planning at L1: no requirement. 5.2.3 Execution. 5.2.3.1 Tool calling: the speaker shall be able to call a determinate tool to complete a single-step task. 5.2.3.2 Content generation: no requirement. 5.2.3.3 Expressive output: the speaker shall be able to convert text content into speech for output. 5.2.3.4 Interconnected collaboration: no requirement. 5.2.4 Memory: no requirement.

5.3 L2 tool level. 5.3.1 Perception. 5.3.1.1 Perception of user information includes the following. a) The speaker shall be able to collect the audio information the user inputs and be woken by it: 1) the success rate of speech wakeup in a low-noise environment, with a sound intensity below 45 dB, shall be not lower than 80 %; 2) the success rate of speech wakeup in a high-noise environment, with a sound intensity from 45 dB to 60 dB, shall be not lower than 65 %; 3) the wakeup time shall be not longer than 1 s. b) The speaker shall be able to recognize the words and sentences in the speech of the user, with a character error rate not higher than 15 % in a low-noise environment and not higher than 20 % in a high-noise environment. c) In particular scenarios the speaker shall be able to carry on the

conversation continuously without the user having to use the wake word again.

5.3.1.2 Perception of device information: the speaker shall be able to monitor basic software and hardware state information, including a) hardware state information such as the volume setting and the state of the network connection; b) software state information such as the running state of the system and the content being played; c) information about externally connected devices, such as their type, capability and running state.

5.4.2.2 Reasoning at L3 includes b) the capability to carry out complex reasoning by combining information of different sources and types, such as multiple constraint conditions, multi-step reasoning or a long chain of thought. 5.4.2.3 Planning includes a) the capability to break a complex task down into executable subtasks; b) the capability to arrange a sensible sequence of execution for the subtasks obtained.

5.4.3 Execution. 5.4.3.1 Tool calling includes a) the capability to call a determinate tool to complete a single-step task; b) the capability to call a determinate combination of tools to complete a preset and clearly defined multi-step task; c) the capability to select and call suitable tools dynamically according to the planning result so as to complete the task.

5.4.3.2 Content generation: the speaker shall be able to generate speech content, and the MOS of the generated result shall be not lower than 4 points. 5.4.3.3 Expressive output: the speaker shall be able to convert text content into speech for output, and the MOS of the synthesized result shall be not lower than 4 points.

5.4.3.4 Interconnected collaboration includes a) the capability to control external devices such as smart home appliances; b) the capability, in a multi-device environment, to select automatically the most suitable device to respond to the instruction when the user utters the wake word; c) the capability to migrate content across devices; d) the capability for several devices to play content in synchrony.

5.4.4 Memory. 5.4.4.1 Short-term memory: the speaker shall be able to remember the content of the context of a single session. 5.4.4.2 Long-term memory: the speaker shall be able to remember content over the long term, such as the basic information of the user, preference settings, the record of interaction history, location habits, diary and tasks, application usage habits, the history of device states, biometric feature templates and the association of behaviour across applications.

6 Determination of the level

Where the speaker possesses all the capabilities of a given level, the speaker is regarded as having reached that level of intelligence. The test methods are given in Annex A; the typical use scenarios used for the test may follow Annex B, and the method of designing scenarios may follow Annex C.