ICS 35.240.01
L 77

中华人民共和国国家标准

GB/T XXXXX.1—XXXX

信息技术　智能语音交互测试方法 第1部分：语音识别

Information technology—Intelligent speech interaction testing method—Part 1: Speech recognition

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

GB/T XXXXX《信息技术　智能语音交互测试方法》为GB/T 36464《信息技术　智能语音交互系统》提供基础性测试方法。

本文件是GB/T XXXXX《信息技术　智能语音交互测试方法》的第1部分。GB/T XXXXX已经发布了以下部分：

——第1部分：语音识别；

——第2部分：语义理解。

本文件由全国信息技术标准化技术委员会（SAC/TC 28）提出并归口。

本文件起草单位：XXX。

本文件主要起草人：XXX。

引　言

智能语音交互在智能家居、智能客服、移动终端、车载终端等诸多领域应用广泛，已成为当前人机交互的重要方式之一。随着智能语音交互越来越深入到生产生活的方方面面，需要对智能语音交互的系统参考框架、基础技术要求、互联网接口要求等进行统一规范，在这方面，我国已制定了支撑智能语音交互系统的基础性国家标准。在此基础上，也需要用统一的测试方法和评价标准来对智能语音交互系统的能力进行评测，为智能语音交互相关的产品和服务提供评测的基础方法和依据。

智能语音交互包括语音识别、语义理解和语音合成三个基本环节，各环节所涉及的测试对象、测试项目、测试环境和测试方法均有所不同。GB/T XXXX《信息技术　智能语音交互测试方法》旨在确立和描述适用于智能语音交互各环节的测试项和测试方法，拟由三个部分构成。

——第 1 部分：语音识别。目的在于为智能语音交互中的语音识别环节提供测试项和测试方法。

——第 2 部分：语义理解。目的在于为智能语音交互中的语义理解环节提供测试项和测试方法。

——第 3 部分：语音合成。目的在于为智能语音交互中的语音合成环节提供测试项和测试方法。

信息技术　智能语音交互测试方法　第1部分：语音识别

1　范围

本文件描述了智能语音交互测试中语音识别（子）系统的测试项和测试方法。

本文件适用于智能语音交互测试中语音识别（子）系统测试的设计和实施。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 21023　中文语音识别系统通用技术规范

GB/T 36464（所有部分）　信息技术　智能语音交互系统

3　术语和定义

GB/T 36464界定的以及下列术语和定义适用于本文件。

3.1

说话人分离　speaker diarization

对包含有效语音信号的音频流中的多个说话人进行说话人分割和说话人聚类的过程。

注：说话人分离的目的是一般是对空间中存在的多个说话人进行分类和追踪。

3.2

说话人分割　speaker segmentation

在多个说话人中找出说话人改变的时间边界，并根据这些边界将音频流切分成多个语音片段。

3.3

说话人聚类　speaker clustering

将属于同一个说话人的一个或多个语音片段进行归类。

3.4

语音编码　speech coding；speech encoding

语音波形编码　speech waveform coding

按照一组能合理重构语音信号的规划，由经数字化的语音信号到离散的数据元序列的转换。

注：语音数字化可与用于语音压缩的某一编码相结合。因此，“语音编码”这一术语常指这种组合运算。

[来源:GB/T 5271.29—2006,29.01.23]

3.5

汉语方言 Chinese dialect

汉语中跟普通话有区别，只在特定地区使用的话。

4 概述

智能语音交互测试中的语音识别测试主要包括功能测试与性能测试，具体如下：

—— 功能测试用于检查被测系统是否提供了语音识别相关的各项功能，具体包括：语音转文字，语音唤醒，前端信号处理，说话人分离，语言信息识别，识别后处理；

—— 性能测项用于检测被测系统中语音识别相关的各项性能，具体包括：识别效果，识别效率，语音唤醒效果，前端信号处理效果，说话人分离效果，语言信息识别效果，系统稳定性。

测试过程中，根据被测系统技术规范进行在线/离线的功能和性能测试。

注：本文件对所列出的功能测试项选择不做要求，测试方根据被测系统的功能要求和测试需求选择测试项。

5 测试准备和执行

5.1 测试数据集

在测试开始前，应通过提前录制或采集的方式制作测试数据集。可根据不同测试项划分出多个测试数据集，在实际测试时可根据需要选择测试数据集。测试数据集类型和要求应符合表1的要求。

表1 测试数据集类型和要求

语音质量	语音种类									
	男声普通话	女声普通话	儿童普通话	老人普通话	男声汉语方言	女声汉语方言	无有效语音内容语音	系统支持的语言信息类型语音	系统支持/不支持的语言信息类型语音	空音频
无噪声正常	A类	A类	A类	A类	B类	B类	C类	C类	C类	C类
弱噪声正常	A类	A类	B类	B类	B类	B类	D类			
强噪声正常	B类	B类	C类	C类	C类	C类	D类			
大音量	B类	B类	B类	B类	D类	D类	D类			
快语速	B类	B类	C类	C类	D类	D类	D类			
截断音频	C类	C类								

测试数据应满足以下要求。

a) 测试语音至少2000条，其中，各类测试语音数量要求如下：

1) A类的总量不得小于测试总量的70%；

2) B类的总量不得小于测试总量的15%，不得大于测试总量的20%；

3) C类的总量不得小于测试总量的5%，不得大于测试总量的10%；

4) D类为可选，总量不得大于测试总量的5%。

b) 各种语音种类的发音人，不得少于30名。

c) 3 s～5 s时长的测试语音应占测试总量的80%以上。

5.2 测试工具

a) 可编程测试工具要求如下：
- 应能调用被测系统开放接口；
- 应能对工具配置文件进行定制；
- 应能接收语音数据并将其输入至被测系统；
- 应能进行功能测试及其相应的性能测试；
- 应能以文本形式获取被测系统运行结果。

b) 测试统计工具要求如下：
- 应能对不同测试参数的系统运行结果进行统计和分析；
- 应能对系统运行结果和标准结果对比文件进行比对。

c) 资源监测工具应能监测内存、CPU、GPU、句柄数等系统资源参数。

5.3 测试设备

音频采样设备：音频采样设备参数应符合表2的要求。

表 2 音频采样设备参数要求

设备名称	参数要求
可移动的声卡	支持 44.1 kHz 及以上的采样频率，16 bit 及以上的模数转换器和数模转换器。
录音软件	波形采样范围为±5000 smpl～±10000 smpl。
计算机	支持录音软件的安装和使用。
声压计	用于环境声压确认。

传声器设备：传声器设备参数应符合表3的要求。

表 3 传声器参数要求

参数名称	符号	测试条件	最小值	典型值	最大值
灵敏度	S/(dBV/P_A)	94 dB SPL@1kHz	-45	-42	-39
信噪比	SNR/(dB(A))	94 dB SPL@1kHz	/	/	/
输出阻抗	Z_{out}/(Ω)	94 dB SPL@1kHz	/	/	400
总谐波失真	THD+N/(%)	100 dB SPL@1kHz	/	/	1
		115 dB SPL@1kHz	/	/	10
指向性	—	全指向性	/	/	/

回放设备：回放设备参数应符合表4的要求。

表 4 回放设备参数要求

设备名称	参数要求	说明
计算机	支持音频播放软件的安装和使用	/
回放外部环境	外界噪声不超过 55 dB(A)情况下，室内本底噪声≤20 dB(A)，（周围无明显振动源，关闭通风） 截止频率 80 Hz	

表 4　回放设备参数要求（续）

设备名称	参数要求	说明
播放器	频率响应（±2.5 dB）：74 Hz～18 kHz 最大声压级：102 dB(A)	可在无人工嘴的条件下使用
功率放大器和人工嘴	信噪比：90 dB 增益控制：0 dB～25 dB 频率响应：200 Hz～10 kHz 最大声压级：110 dB(A)	具体场景按照 GB/T 36464 执行
噪声播放音箱	功率：70 W （峰值 125 W） 频响：50 Hz～21 kHz 声压：≤113dBSPL@1m（对） 输入阻抗：10KΩ 最大输入电平：22 dBu	
仿真人体	根据播放器和人工嘴的尺寸和安装位置定制	/

5.4　测试环境

应根据被测系统的特定功能和性能要求以及使用场景配置相应的软硬件环境。

5.5　测试执行

应使用可编程测试工具和测试统计工具将测试数据集输入到在线/离线状态的被测系统中并获取运行结果。

5.6　结果判定

应根据被测系统技术要求对系统在各测试项上的运行结果进行记录、分析和判定，形成测试结果。

6　功能测试方法

6.1　语音转文字

测试内容：检查被测系统是否提供将所接收到的有效语音信号转化为文字，并输出与语音内容相符的结果的功能。

测试方法：按照表1的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.2　语音唤醒

测试内容：检查被测系统是否提供语音唤醒的功能，包括但不限于以下具体功能。

a)　命令字（词）唤醒：能使用预定义的唤醒命令字（词）唤醒被测系统。
b)　自定义唤醒命令字：能自定义唤醒命令字（词）。
c)　多命令字唤醒：能使用不同的唤醒命令字（词）唤醒被测系统。
d)　多音频流监听：被测系统在执行语音唤醒的同时能监听多个音频流。
e)　语音打断唤醒：能使用语音打断的方式唤醒被测系统。

f) 协同唤醒：使用相同命令字（词）的多个设备在同一场景中出现，一次唤醒操作有且仅有一个设备应答。

测试方法：按照表1的要求制作包含预定义唤醒命令字、非唤醒命令字、自定义唤醒命令字、多个唤醒命令字和语音打断唤醒命令字的测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.3 前端信号处理

测试内容：检查被测系统是否提供前端信号处理的功能，包括但不限于以下具体功能。

a) 语音编解码：被测系统能设置语音编码算法的压缩等级，能支持不同音频编码格式下的语音音频的压缩和解压缩，并确保语音内容不变。
b) 端点检测：被测系统能从连续音频流中检测出第一个或多个语音片段的起始点和结束点，能设置端点检测灵敏度，即设置语音等待超时时长和尾部静音长度。
c) 语音增强：被测系统能自动对输入语音的信噪比进行改善，能对输入语音中的背景噪声和晚期混响进行抑制。
d) 声源定位：被测系统能自动对发声源的空间位置进行定位。
e) 格式转换：被测系统能自动对输入音频的格式进行转换，并确保语音内容不变。
f) 重采样：被测系统能改变数字语音信号的采样率，并确保语音内容不变。
g) 音频质量判断：检查被测系统是否提供对输入音频质量进行自动判断的功能。
   示例：对音量过小、信噪比过低或存在前、后截断的音频判断为音频质量较差。
h) 声学回声消除：检查被测系统是否提供对输入音频进行回声消除的功能。

测试方法：按照表1的要求制作包含多种音频质量的测试数据集，包括前截断音频、后截断音频、音量小音频、信噪比低音频等，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.4 说话人分离

测试内容：检查被测系统是否提供说话人分离的功能，包括但不限于以下具体功能。

a) 说话人分割：被测系统能自动进行说话人分割，分割后的语音片段只包含一个说话人的语音内容。
b) 说话人聚类：被测系统能自动进行说话人聚类，聚类后的语音片段分别对应不同的说话人。

测试方法：按照表1的要求制作包含多个说话人交替对话的测试数据集，对话时长宜20 min，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.5 语言信息识别

测试内容：检查被测系统是否提供语言信息识别的功能，包括但不限于以下具体功能。

a) （多）语种识别：被测系统能自动对输入语音所属的某个或多个（2 个及以上）语种进行判断并输出（多）语种识别结果。
b) 多语种混读识别：在多个语种混读的情况下，被测系统能自动对不同语种进行判断并输出多语种混读识别结果。
c) （多）汉语方言识别：被测系统能自动对输入语音所属的某个或多个（2 个及以上）汉语方言进行判断并输出（多）汉语方言结果。
d) 语言信息端点识别：被测系统能自动对不同语种、汉语方言的音频片段端点进行判断并输出端点识别结果。

测试方法：按照表1的要求制作包含一个或多个语音信息的测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

6.6 识别后处理

测试内容：检查被测系统是否提供对语音识别结果进行后处理的功能，包括但不限于以下具体功能。

a) 数字规整：按照需求将汉字表示的数字规整为符合阅读习惯的格式。

示例：对“一千两百三十二元”进行数字规整后的结果为“1232元”。

b) 字母规整：按照需求将识别结果规整为更易阅读的格式。

示例：对“三W点”进行字母规整后的结果为“www.”。

c) 繁简体规整：按照需求将识别结果规整为繁体或简体。

示例：对“一二三四五”进行繁简体规整后的结果为“壹贰叁肆伍”。

d) 标点预测：根据语音内容对识别结果添加中文或者英文标点。

示例：对“你好你在哪里”进行标点预测后的结果为“你好，你在哪里？”。

e) 文本顺滑：根据语音内容过滤识别结果文本中无意义的字或词。

示例：对“额……额……我不太清楚”进行文本顺滑后的结果为“我不太清楚”。

f) 文本替换：根据语音内容动态替换或修改识别结果文本中的某些字或词。

示例：对“它们做出了很大的牺牲”进行文本替换后的结果为“他们作出了很大的牺牲”。

测试方法：按照表1的要求制作测试数据集，使用可编程测试工具和测试统计工具将测试数据集输入到被测系统并获取运行结果，按照测试内容的描述对结果进行判定。

7 性能测试方法

7.1 识别效果

测试内容：识别效果测试项包括对字识别效果和句识别正确率的测试。

a) **字识别效果**：字识别效果由字匹配率等参数表征，它们共同显示被测系统的字识别能力。参数值按GB/T 21023描述的方法计算：
   - 字匹配率，
   - 替代错误率，
   - 插入错误率，
   - 删除错误率，
   - 字错误率，
   - 字准确率。

b) **句识别正确率**：此参数显示被测系统的句识别能力，参数值计算方法为被测系统正确识别的句子数量除以标注的总句子数量。

测试方法：识别效果测试方法如下。

—— 测试数据：按照表1的要求制作测试数据集，对各测试数据集所有的语音内容进行人工标注，并制作成标准结果对比文件。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求。

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.2 识别效率

测试内容：识别效率测试项用识别平均响应时间、识别平均子句响应时间、识别响应时间分布和识别吞吐率等参数表征。

a) **识别平均响应时间**：识别响应时间指的是接收到一条语音后，被测系统给出该条语音识别结果的时间；识别平均响应时间是测试数据集上所有识别响应时间与输入语音总条数的比值。参数值计算方法见公式（2）：

$$T_{ack}^{avg} = \sum_{i=1}^{N} (t_r^i - t_e^i)/N \quad \text{（式2）}$$

式中：

T^{avg}_{ack}——识别平均响应时间；

t^i_r ——得到第i条语音识别结果的时刻；

t^i_e ——第i条语音输入结束的时刻；

N ——输入语音总条数。

b) **识别平均子句响应时间**：识别子句响应时间指的是接收到一条语音后，被测系统给出该条语音中某一子句识别结果的时间；识别平均子句响应时间是测试数据集上所有识别子句响应时间与输入语音总条数的比值。参数值计算方法见公式（3）：

$$T_{scl}^{avg} = \sum_{i=1,j=1}^{N} (t_r^{ij} - t_e^{ij})/N \quad \text{（式3）}$$

T^{avg}_{scl} ——识别平均子句响应时间；

t^{ij}_r ——得到第i条语音中第j个子句识别结果的时刻；

t^{ij}_e ——第i条语音中第j个子句输入结束的时刻；

N ——输入语音总条数。

c) **响应时间分布**：此参数显示测试数据集上所有的响应时间的分布及其占比情况。

d) **识别吞吐率**：即被测系统在单位响应时间内识别语音音频的时间长度。参数值计算方法为测试集上语音音频总时长除以测试集上响应时间总时长。

测试方法：识别效率测试方法如下。

—— 测试数据：按照表1的要求制作测试数据集，此外，制作语音时长为10 s±0.1 s、语音结束后无静音的测试数据集用于平均响应时间和响应时间分布测试；制作语音时长大于10小时的测试数据集用于识别吞吐率测试。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求。

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.3 语音唤醒效果

测试内容：语音唤醒效果测试项包括唤醒率，误唤醒频度，语音打断成功率、语音打断唤醒率、语音打断误唤醒频度、协同唤醒成功率和协同唤醒正确率等参数，具体参数描述如下。

a) **唤醒率：**检测被测系统的语音唤醒能力，即某段时间内的语音唤醒操作中，成功唤醒被测系统的比率。参数值计算方法为成功语音唤醒次数除以语音唤醒操作总次数。

b) **误唤醒频度：**检测被测系统的语音唤醒能力，即单位时间内被测系统被误唤醒的次数。参数值计算方法为误唤醒次数除以测试总时长。

c) **语音打断成功率：**检测被测系统的语音打断能力，即被测系统成功被语音打断的比率。参数值计算方法为语音打断成功次数除以语音打断操作总次数。

d) **语音打断唤醒率：**检测被测系统的语音唤醒能力，即某段时间内的语音唤醒操作中，成功打断并唤醒被测系统的比率。参数值计算方法为成功打断唤醒次数除以语音打断唤醒操作总次数。

e) **语音打断误唤醒频度：**检测被测系统的语音唤醒能力，即单位时间内被测系统语音打断误唤醒的次数。参数计算方法为语音打断误唤醒次数除以测试总时长。

f) **协同唤醒成功率：**检测同一场景多个设备的语音唤醒能力，即对多个设备进行唤醒操作，有且仅有一个设备被成功唤醒的比率。参数计算方法为有且仅有一个设备被成功唤醒的次数除以语音唤醒操作总次数。

g) **协同唤醒正确率：**检测同一场景多个设备的语音唤醒能力，即对多个设备进行唤醒操作，用户意图中的设备被正确唤醒的比率。参数计算方法为用户意图中的设备被正确唤醒的次数除以语音唤醒操作总次数。

测试方法：语音唤醒效果测试方法如下。

—— 测试数据：按照表1的要求制作包含预定义唤醒命令字、非唤醒命令字、自定义唤醒命令字、多个唤醒命令字和语音打断唤醒命令字的测试数据集，宜选取不少于200条测试语音。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求，其中，测试场景类型见表5。

表 5 语音唤醒效果测试场景

场景编号	场景描述	信噪比 dB
场景 1	安静场景	＞25
场景 2	低噪场景	15～25
场景 3	高噪场景	0～15
场景 4	自定义噪音场景	具体场景按照 GB/T 36464 执行

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.4 前端信号处理效果

测试内容：前端信号处理效果测试项包括语音编解压缩率、端点检测准确率、信噪比改善率、噪声抑制量、声源定位正确率和回声返回衰减增益等参数，具体参数描述如下。

a) **语音编解压缩率**：检测被测系统的语音编解码能力，参数计算方法为语音编码后输出的压缩音频码流率除以语音编码前音频码流率。

b) **端点检测准确率**：检测被测系统的端点检测能力，参数计算方法见公式（13）：

$$A_{VAD} = \frac{(T_{valid} + T_{silence})}{(T_{valid} + T_{silence} + T_{dev})} \times 100\% \quad \text{（式13）}$$

式中：

A_{VAD} ——端点检测准确率；

T_{valid} ——有效语音音频时长；

$T_{silence}$——静音音频时长；

T_{dev} ——检测误差音频时长。

c) **信噪比改善率**：检测被测系统的语音增强能力，参数计算方法为输出语音的信噪比除以输入语音的信噪比。

d) **噪声抑制量**：检测被测系统的语音增强能力，即被测系统输出信号的噪声振幅相对于输入信号的噪声振幅的减少量。参数计算方法见公式（15）：

$$D_{NR} = 10\log\sum_{n=0}^{N-1}|v_{in}(n)|^2 \Big/ \sum_{n=0}^{N-1}|v_{out}(n)|^2 \quad \text{（式15）}$$

式中：

D_{NR} ——噪声抑制量；

N ——输入信号频谱频率分量的总数量

$v_{in}(n)$ ——输入信号中第n个噪声信号的振幅；

$v_{out}(n)$——输出信号中第n个噪声信号的振幅。

e) **声源定位正确率**：检测被测系统的声源定位能力，参数计算方法为声源定位正确次数除以声源定位请求总次数。

f) **回声返回衰减增益**：检测被测系统的回声消除能力，参数计算方法见公式（17）：

$$ERLE = 10\log_{10}(E\{|y(n)|^2\}/E\{|e(n)|^2\}) \quad \text{（式17）}$$

式中：

$ERLE$——回声返回衰减增益，单位dB；

$y(n)$——期望回声信号；

$e(n)$——误差信号。

测试方法：前端信号处理效果测试方法如下。

—— 测试数据：按照表1的要求制作测试数据集，此外，制作前、后静音段时长不少于3秒的测试数据集，宜不少于200条测试语音，用于端点检测准确率测试。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求。

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.5 说话人分离效果

测试内容：说话人分离效果测试项包括分离错误率参数：检测被测系统的说话人分离能力，即被测系统分离错误的语音片段时长占整个有效语音片段时长的比率。参数计算方法见公式（18）：

$$DER = \frac{\sum_{s=1}^{S} dur(s) \times \left(max\left(N_{ref}(s), N_{hyp}(s)\right) - N_{correct}(s) \right)}{\sum_{s=1}^{S} dur(s) \times N_{ref}} \quad \text{（式18）}$$

式中：

DER ——分离错误率；

S ——实际结果和系统输出结果都包含同一个说话人（对）的说话人片段数量；

$dur(s)$ ——片段s的时长；

$N_{ref}(s)$ ——片段s中实际结果的数量；

$N_{hyp}(s)$ ——片段s中系统输出结果的数量；

$N_{correct}(s)$——片段s中系统输出结果与实际结果正确对应的数量。

测试方法：说话人分离效果测试方法如下。

—— 测试数据：按照表1的要求制作包含至少2个说话人交替对话的测试数据集，对话时长宜20 min。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求。

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.6 语言信息识别效果

测试内容：语言信息识别效果测试项包括召回率参数，参数计算方法为某类型语言信息被正确识别的次数除以其应该被识别正确的次数。

测试方法：语言信息识别效果测试方法如下。

—— 测试数据：按照表1的要求制作包含一个或多个语音信息的测试数据集，选取不少于200条测试语音。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求，其中，测试场景类型见表5。

—— 测试执行：按照5.5的要求对被测系统进行测试。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

7.7 系统稳定性

测试内容：系统稳定性测试项包括稳定运行和资源使用等参数。

a) **稳定运行**：检测在给定的软硬件配置和系统并发路数的条件下，检测被测系统运行 6.1～6.5 的各项功能，未出现崩溃、假死或功能异常，且符合性能要求，能持续正常运行的能力。

注：给定的软硬件配置和系统并发路数需满足被测系统正常运行的能力。

b) **资源使用**：检测在给定软硬件配置和系统并发路数的条件下，检测被测系统运行 6.1～6.5 的各项功能，系统磁盘性能、内存、CPU、网络资源等各项资源使用率持续平稳的能力。

注：给定的软硬件配置和系统并发路数需满足被测系统正常运行的能力。

测试方法：系统稳定性测试方法如下。

—— 测试数据：按照测试项要求准备测试数据集，并明确软硬件配置和系统并发路数。

—— 测试工具：符合5.2的要求。

—— 测试设备：符合5.3的要求。

—— 测试环境：符合5.4的要求。

—— 测试执行：按照5.5的要求对被测系统进行测试，在线场景下持续7天、离线场景下持续3天连续不间断向被测系统循环输入测试语音，连续监测系统运行情况和系统磁盘性能、内存、CPU、网络资源等各项资源使用率变化情况。

—— 结果判定：按照测试内容描述的方法得出系统运行结果并生成结果文件，包括测试数据集名称、测试数据集数量、软硬件配置、系统并发路数和测试项结果等。系统运行结果符合被测系统技术要求或相关标准规范则测试通过，否则不通过。

参 考 文 献

[1] GB/T 5271.29—2006 信息技术 词汇 第29部分：人工智能 语音识别与合成
[2] GB/T 21024—2007 中文语音合成系统通用技术规范