

T/XXX

团 体 标 准

T/XXXX—XXXX

粤语智能语音服务系统技术要求

Specification of Cantonese intelligent customer service speech interaction system

(征求意见稿)

XXXX – XX – XX 发布

XXXX – XX – XX 实施

XXXXXX 发布

目次

前言..... III

引言..... IV

1 范围..... 1

2 规范性引用文件..... 1

3 术语和定义..... 1

4 系统功能要求..... 1

 4.1 系统功能构成..... 1

 4.2 语音识别服务..... 1

 4.3 语音合成服务..... 2

 4.4 文本处理服务..... 2

 4.5 服务支撑..... 2

 4.6 服务管理..... 2

 4.7 运行监控..... 2

5 系统性能要求..... 3

 5.1 概述..... 3

 5.2 性能评价项..... 3

 5.3 性能评价要求..... 4

6 系统服务要求..... 5

 6.1 系统服务流程..... 5

 6.2 系统数据输入..... 5

 6.3 系统服务输出..... 5

7 测试方法..... 5

 7.1 概述..... 5

 7.2 测试数据准备..... 6

 7.3 测试方法..... 7

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。

本文件由广东省标准化研究院提出。

本文件由 XXXXX 归口。

本文件起草单位：广东省标准化研究院。

本文件主要起草人：王佳胜、谢子紫。

引 言

近年来，我省深入实施创新驱动发展战略，语音质检普遍被应用在智能外呼和客服领域，通过语音识别与声纹识别的相关技术，不仅可以对客户说话的内容进行语音语义分析，挖掘客户潜在需求，进行用户画像，提供个性化的客户服务与产品的精准营销，还可以对对话内容的合规性进行稽核与审查，进一步提升服务满意度。为规范粤语智能语音服务系统，特制定本文件。

粤语智能语音服务系统技术要求

1 范围

本文件规定了粤语智能语音服务系统的功能要求、性能要求以及服务要求。

本文件适用于使用粤语的智能客服领域及相关业务平台实现粤语智能语音服务系统的设计、开发、应用、测试和维护。

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 5271.29-2006 信息技术 词汇 第29部分：人工智能 语音识别与合成

GB/T 21023-2007 中文语音识别系统通用技术规范

GB/T 21024-2007 中文语音合成系统通用技术规范

GB/T 36464.3-2018 信息技术 智能语音交互系统 第3部分：智能客服

3 术语和定义

GB/T 5271.29-2006以及GB/T 21024-2007界定的术语和定义适用于本文件。

4 系统功能要求

4.1 系统功能构成

系统应包含语音识别、语音合成、文本处理等服务组件，并具备对系统服务支撑、管理和运行监控等功能。

4.2 语音识别服务

4.2.1 一般要求

语音服务应能实现下列功能：

- 关键词语音识别；
- 连续语音识别；
- 数字识别；
- 多语种语音识别，一般包括汉语、英语；
- 方言的确差异语音识别；
- 针对特定方言进行模型定制；
- 对待识别音频的有效语音进行前后端点检测，从连续音频流分离出有效语音；
- 对识别的音频进行干扰声抑制，提高语音识别准确率；
- 说话人识别；
- 混合语言识别。

4.2.2 专业要求

除一般要求外，语音识别服务还应依据不同地区的实际需求，提供以下功能：

- 自定义热词，并在语音识别中生效；
- 通用要素词汇的语音识别优化和个性化要素词汇的语音识别拓展；
- 对呼叫中心的录音或实时对话内容进行语音识别，帮助呼叫中心快速发现服务中的问题；
- 根据实时语音识别内容进行客群画像。

4.3 语音合成服务

语音合成服务应能实现下列功能：

- 多语种语音合成，一般包括汉语、英语等；
- 多方言语音合成，针对广东不同地区粤语特点进行语音合成；
- 多音色语音合成，一般包括青年男声、青年女声等，支持定制个性化音色语音合成。
- 按照指定的文本分词方式或发音方式进行语音合成。

4.4 文本处理服务

文本处理服务应能实现下列功能：

- 从文本中正确处理简称、别称、代码、数字以及典型错别词；
- 对符合预设格式的文本语义进行抽取和结构化处理，包括但不限于概念间语义理解（上下位关系理解）、部分整体关系理解、地理位置关系理解、因果关系理解、时序关系理解、近义关系理解；
- 从文本中提取服务要素，对照相应的标准化要素信息库进行比对修正。

4.5 服务支撑

系统服务支撑应能实现下列功能：

- 可支持电话、网站、app 等全渠道平台覆盖；
- 支持用户体验各类服务的应用效果；
- 实现客服队列调度及坐席分配，基于策略模型做到智能分流与智能分发；
- 为人工坐席提供话术推荐、业务指导与应答提示等功能。

4.6 服务管理

系统服务管理应能实现下列功能：

- 对用户进行管理，支持用户信息新增、修改和删除等操作；
- 对服务进行分类授权，支持限制其访问服务类型、次数、使用期限；
- 对服务进行认证鉴权，符合用户身份认证体系要求。

4.7 运行监控

- 支持对各服务组件全生命周期的运行状态进行监控管理；
- 支持对服务次数、服务时间、服务高峰期、用户使用总量等指标进行统计；
- 对系统服务的录音或实时对话内容进行语音分析，帮助快速发现服务中的问题；
- 根据业务需求人工增加、删除、修改客服系统服务的提示用语；
- 提炼系统服务过程中的高绩效优质话术，实现服务营销策略推荐，帮助客服销售人员复盘，对客户沟通相关的话术、方式及策略给予指导。

5 系统性能要求

5.1 概述

语音识别性能主要通过字准确率、数字语音识别正确率、语音端点检测正确率、语义理解正确率、语音交互成功率和语音识别平均响应时间等指标进行评价；语音合成性能主要通过语音合成质量进行评价。

5.2 性能评价项

5.2.1 字准确率

字准确率应符合以下要求：

- 在低噪环境（噪声强度在 50 dB 以下）中，关键词语音识别的字正确率宜在 85% 以上；在高噪环境（噪声强度在 60 dB~65 dB）中，关键词语音识别的字正确率宜在 85% 以上。
- 在低噪环境（噪声强度在 50 dB 以下）中，连续语音识别的字正确率宜在 85% 以上；在高噪环境（噪声强度在 60 dB~65 dB）中，连续语音识别的字正确率宜在 85% 以上。

5.2.2 数字语音识别正确率

数字语音识别正确率应在99%以上。

5.2.3 语音端点检测正确率

语音端点检测正确率应为一段音频中每一帧被标注为语音帧或静音帧的正确率，当语音帧或静音帧的标准结果和判断结果相一致时，则该帧分类正确。

语音端点检测正确率的计算方法见式（1）。

$$CR_{VAD} = C_{VAD}/S_{VAD} \times 100 \% \cdots \cdots \cdots (1)$$

式中：

- CR_{VAD} ——语音端点检测正确率；
- C_{VAD} ——对于该段音频，分类正确的帧数；
- S_{VAD} ——音频总帧数。

5.2.4 语义理解正确率

语义理解能力经过不同知识背景、不同年龄的多人、多轮次测试，客观测试准确率应达到95%以上，主观测试准确率应达到85%以上。语义理解正确率计算方式见式（2）。

$$R_{SS} = \frac{N_{SS}}{N} \times 100 \% \cdots \cdots \cdots (2)$$

式中：

- R_{SS} ——语义理解正确率， %；
- N_{SS} ——操作意图及语义要素均被正确判断的次数；
- N ——输入被正确识别出文本信息的总次数，或输入正确文本的总次数。

5.2.5 语音交互成功率

系统的交互过程中，用户能够在既定交互轮次内获得目标结果，即可判定为本轮交互成功；反之则判定为本轮交互失败。语音交互成功率应在80%以上。
交互成功率的计算方法见式（3）。

$$P_S = \frac{S}{S+F} \times 100 \% \cdots \cdots \cdots (3)$$

式中：
P_s——交互成功率；
S——交互成功的次数；
F——交互失败的次数。

5.2.6 语音识别平均响应时间

语音识别平均响应时间的计算应符合GB/T 21023—2007中5.3的要求。

5.2.7 语音合成质量

语音合成质量通过对合成语音同真人语音在音质和自然度等方面的差异进行主观评测，采用平均意见得分进行计算。主观评测标准见表1。

表1 主管评测标准

评分	测定效果
5	优秀，语音自然，音色接近真实
4	较好，仅能察觉若干不自然
3	尚可，能察觉到不自然但可以接受，评分分界线
2	较差，明显察觉不自然并不愿意接受
1	极差，无法接受

5.3 性能评价要求

按照5.2给出的评价项，根据各业务需要，挑选不低于1000条语料形成测试集，对系统性能进行评价。系统性能指标要求见表2。

表2 平台性能评价要求

	评价项						
	语言类型						

普通话							
-----	--	--	--	--	--	--	--

表 2（续）

粤语							
英语							

6 系统服务要求

6.1 系统服务流程

各部门根据自身需要调用系统，实现语音识别、语音合成和文本处理等功能。

6.2 系统数据输入

在系统输入的数据包括语音数据和文本数据，其中：

- 语音数据：用户在语音采集端以说话的形式输入语音；
- 文本数据：用户在系统上输入的文本资料。

6.3 系统服务输出

系统将语音识别、语音合成和文本处理结果返回系统储存。

7 测试方法

7.1 概述

测试方法的评价方式如下：

a) 系统测试分为机器测试和人工测试，其中：

- 1) 机器测试指应用机器系统, 利用事先准备好的测试数据进行系统测试。
- 2) 人工测试指由不同知识水平、不同年龄阶段的测试人员完成的人工测试。

b) 评价方式分为客观测试方式和观测试方式, 其中：

1) 客观测试方式, 指事先定义好问题列表 (可以是语音表达的问题，也可以是文字表达的问题)，并已经知道每个问题对应的答案。测试样本所涉及的语法、语料经过后台系统学习后的测试。

2) 主观测试方式, 指事先不进行问题列表定义，由测试者根据对测试场景、目的主观理解进行测试，目的是测试系统功能是否适应不同学历背景、不同年龄段的人群使用。测试结果由测试者主观进行评价。

可以针对智能客服语音交互系统分模块进行测试,如单独进行语音识别、语音合成 语义理解、智能搜索、维护能力测试;也可以针对人机交互全流程进行测试。即不单独考虑语音识别或语义理解的准确率,只判别人机交互准确成功的测试。

7.2 测试数据准备

7.2.1 测试数据分类

测试数据分为两类:

a) 文本类型测试数据,指覆盖智能客服服务范围的文本类型信息的测试数据,包含关键词、问题文本和答案等。

b) 语音类型测试数据,指针对覆盖智能客服服务范围的文本类型信息的测试数据,选取不同性别、年龄、知识背景的测试人员正常发音的中文普通话语音数据,包含关键词、问题文本的语音数据。

7.2.2 测试数据要求

测试数据要求如下:

a) 为验证智能客服的服务能力、智能水平、系统的易用性和可维护性,应事先针对服务内容和领域定义文本形式的问题、对应答案等;测试语料应从客服系统服务内容的词汇量覆盖、已定义业务覆盖、开放业务覆盖以及常用性角度加以设计;

b) 测试语料宜覆盖针对数字及字母的语料;

c) 对于关键词识别的系统,测试语料应覆盖典型业务场景的标准词汇,宜覆盖服务业务的所有词汇,测试语料规模应不少于 500 句;

d) 为实现系统测试的可再现性,比较不同系统之间的服务能力,宜提前录制准备测试交互场景定义、对应语音数据(包含对应识别结果标准数据、理解搜索结果标准数据)和环境噪声数据,形成测试标准库,且测试语料、发音人以及噪音环境等设计应保证与实际应用场景的一致性;

e) 智能客服语音交互的测试数据应具备对人机交互系统定义的垂直类问题进行测试的音频、背景噪音音频、文本问题列表和答案列表等;

f) 对于连续语音识别的智能客服语音交互系统-测试语料应覆盖被测系统的常用词汇量,应从常用性角度挑选典型语料,每种开放业务的测试语料规模应不少于500句;

g) 测试录音的说话人的选择应在符合系统对交互人群要求的前提下,选择具有代表性和统计分布规律的发音人,应考虑中文普通话、不同年龄、不同性别、不同语速的测试,可考虑不同口音、不同教育背景和不同说话韵律等因素;

h) 录音人数应不低于60人,应至少分老、中、青三个年龄组,分别为60岁以上、20岁~60岁、20岁以下,进行录音及人工测试。各发音人应分别录制全部的测试语料。

7.2.3 录音过程要求

录制过程应包括录音、标注和确认三个步骤。

7.2.4 录音设备要求

录音设备应符合GB/T 36464.3-2018《信息技术智能语音交互系统第3部分:智能客服》6.2.4的要求。

7.2.5 环境噪声数据录制

环境噪声应符合GB/T 36464.3-2018《信息技术智能语音交互系统第3部分:智能客服》6.2.5的要求。

7.3 测试方法

7.3.1 概述

应采用机器和人工的方式,分别按照测试语料要求进行客观测试,判别系统返回结果是否与事先定义的结果一致。

可采用机器和人工的方法,按照定义的场景进行主观测试,综合两种测试方法的结果对被测系统进行主观评价。

7.3.2 不同语音交互环境的测试

不同交互环境的测试应按如下要求:

- a) 分为实验室语音背景下的测试,实际环境测试,在实际环境测试中,需记录测试环境情况,并用测量仪器测量当时环境噪音强度、信噪比;
- b) 在一定时间内记录返回语音服务结果,记录当次语音交互会话是否成功和有效;
- c) 分别在表2三种噪声环境下按上述a)、b)步骤完成所有语料的测试,并按式(3)分别计算三种噪声环境下的语音交互成功率(保留两位小数点,%)。

7.3.3 语音识别测试

在7.2.5规定的测试环境场景下,使用回放设备播放语音识别测试语料,或人工进行测试语料的输入,记录低噪环境(噪声强度在50 dB 以下)及高噪环境(噪声强度在60 dB~65 dB)下智能客服的识别结果,与预期结果进行比对,统计结果并给出正确率。使用以上测试方法,测试验证被测系统是否满足5.2.1以及5.2.2的要求。

7.3.4 语音合成测试

语音合成测试采用平均意见得分的测试方式进行评价。选取30个体验人员,男女各15人,通过对智能客服语音合成结果的反馈,测听合成语音同真人语音在音质、易懂度和自然度等方面的差异,并以平均意见得分对主观测评进行数值量化,记录平均结果。

使用以上测试方法,测试验证被测系统是否满足4.3的要求。

7.3.5 语义理解测试

采用机器及人工的方式输入文本数据,统计系统反馈的文本的信息。判别正确的次数。测试验证被测系统是否满足5.2.4的要求。

7.3.6 测试结论

通过测试得出各指标测试数据,被测系统测试数据如果通过符合性标准,应评判为可发布使用,否则,应评判为不可发布使用。

测试报告应包括下述指标项的测试结果:

- a) 字准确率;
- b) 数字语音识别正确率;
- c) 语音端点检测正确率;
- d) 语义理解正确率;
- e) 语音交互成功率;
- f) 语音识别平均响应时间;

g) 语音合成质量。
