

语言文字规范

GF 2006 —

文语转换与语音识别系统评测规范

(Assessment Specifications of TTS and ASR System)

2006— — 发布

2006— — 试行

中华人民共和国教育部 国家语言文字工作委员会 发布

目 录

前 言

1	适用范围.....	(4)
2	规范性引用文件.....	(4)
3	术语和定义.....	(5)
4	评测的一般原则和方法.....	(5)
5	文语转换系统的评测.....	(6)
5.1	语音学模块评测.....	(6)
5.2	语言学模块评测.....	(9)
6	语音识别系统的评测.....	(13)
6.1	测试样本和发音人的选择.....	(13)
6.2	语音文本转换测试与音节测试.....	(13)
6.3	语音识别的评测指标.....	(13)

前 言

本标准规定了文语转换和语音识别系统的评测规范。

本标准由教育部语言文字信息管理司提出立项，负责解释。

本标准由教育部语言文字信息管理司归口。

本标准由国家语言文字工作委员会语言文字规范（标准）审定委员会审定。

本标准起草单位：教育部语言文字应用研究所

本标准主要起草人：冯志伟、肖航、富丽、章云帆

文语转换与语音识别系统评测规范

(Assessment Specifications of TTS and ASR Systems)

1 适用范围

本标准规定了文语转换和语音识别系统的评测规范。

本标准适用于文语转换和语音识别系统的评测以及有关的管理工作。

2 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准中的条款。

ISO11179-3 信息技术 数据元的规范与标准化 第3部分：数据元的基本属性

GB/T 12200.1 汉语信息处理词汇 01部分：基本术语

GB/T 13725 信息处理用现代汉语分词规范

GB/T 16403-1996 声学 测听方法 纯音气导和骨听阈基本测听法，eqv ISO 8253-1:1989

GB/T 17696-1999 声学 测听方法 第3部分：语言测听，eqv ISO 8253-3:1996

GB/T 16404.1-1996 声学 声强法测定噪声源的声功率级 第1部分：离散点上的测量，eqv ISO 9614-1:1993

GB/T 16404.2-1999 声学 声强法测定噪声源的声功率级 第2部分：扫描测量 (Acoustics-Audiometric test methods-Part 3: Speech audiometry)，eqv ISO 9614-2:1996

GB/T 13504-1992 汉语清晰度诊断押韵测试(DRT)法(Diagnostic rhyme test (DRT) method of Chinese articulation)

GB/T 16532-1996 通信设备清晰度 DRT 法评价用语音材料库 (Speech material library used DRT for articulation evaluation of communication equipments)，idt IEC 728-1:1986

GB/T 17147-1997 声音广播中音频噪声电平测量，eqv ITU-R 468-4:1990

GB/T 17576-1998 CD 数字音频系统，idt IEC 908:1987

GB/T 14476-1993 客观评价厅堂语言可懂度的 RASTI 法，neq IEC 268-16

GB/T 14919-1994 数字声音信号源编码技术规范 (The specifications for digital audio source coding)，eqv CCIR 646

GB/T 16463-1996 广播节目声音质量主观评价方法和技术指标要求

SJ 20771-2000 军用通讯系统音质 MOS 评价法

GB3259—92 中文书刊名称汉语拼音拼写法

GB/T 15834—1995 标点符号用法

GB/T 15835—1995 出版物上数字用法的规定

GB/T 16159—1996 汉语拼音正词法基本规则

第一批异形词整理表
第一批异体字整理表
部分计量单位名称统一用字表
中国人名汉语拼音字母拼写法
中国地名汉语拼音字母拼写规则
普通话异读词审音表

3 术语和定义

下列术语和定义适用于本规范。

3.1 信息处理系统 (information processing systems)

基于计算机技术、网络互联技术、现代通讯技术和各种软件技术，集各种理论和方法于一体，提供信息服务的人机系统，是由人和计算机等共同组成的，能进行信息的收集、传输、分析、加工、处理、存储、更新和维护的系统。

3.2 评测规范 (assessment specifications)

用于评测的规范，包括评测的一般原则、评测内容、评价指标、评测方法和文件格式等。

3.3 自然语言 (natural language)

规则是根据当前用法而不是用显式的方式规定的语言。如汉语、英语、德语等。

3.5 中文信息处理 (Chinese information processing, CIP)

用计算机对汉语的语音、语法、词汇、语义、语用等信息进行处理，又称汉语信息处理。

3.6 语言文字评测 (language norms based assessment)

依据语言文字的技术指标体系和有关规范，采用一定的方法和程序，对于自然语言信息处理系统及其组成要素中与语言文字相关的部分进行评价和检测。语言文字评测是系统评测的重要组成部分。

3.7 语音合成 (automatic speech synthesis, ASS)

利用语音库和语音合成技术，自动地合成自然语言的语音流。

3.8 文语转换 (text to speech transfer, TTS)

把自然语言的文本自动地转换成语音流。语音合成是文语转换的重要组成部分。

3.9 语音识别 (automatic speech recognition, ASR)

利用语音分析技术，抽取语音特征，自动地把自然语言的语音转换成文本。

3.10 词错误率 (word error rate)

语音识别系统返回的单词串与标准答案中的正确单词串的差别的比率。

4 评测的一般原则和方法

4.1 评测的原则

自然语言信息处理系统的评测应当遵循如下原则：

4.1.1 公平公正的原则：评测应努力作到公平、公正；

4.1.2 遵循标准的原则：评测应遵循国际标准、国家标准和相关语言文字规范；

4.1.3 人机结合的原则：在当前条件下，基本上应当以人工评测为主，辅之以机器自动评测，做到人工评测与机器评测相结合；

4.1.4 区别对待的原则：评测应针对不同语言信息处理系统和用户类型的特点，区别对待，以突出不同系统和用户类型的特殊要求；

4.1.5 灵活柔性的原则：语言文字具有一定的灵活性和柔性，并非处处都是界限分明、非此即彼的，在遵循标准的前提下，有时可以容许两种或多种可能的结果并存。

4.1.6 可操作性的原则：评测应当是可以操作的，评测时，应当具体地说明评测的方法、步骤和评分方式。

4.2 评测的方法

4.2.1 黑箱评测 (black box assessment): 评测时不关心自然语言信息处理系统的内部机制和组成结构, 主要根据系统的输入数据与输出结果进行判断。黑箱评测有助于了解信息处理系统外在的总体性能, 又叫做“外在评测”(extrinsic assessment)。

4.2.2 白箱评测 (glass box assessment): 评测时需要对于自然语言信息处理系统的内部机制分别进行分析, 逐一评测系统的各个组成部分的性能。白箱评测可以针对信息处理系统的各个组成部分分别进行, 对于不同的部分准备不同的测试数据, 从而判断所出现的错误是在那一个部分造成的, 这样就可以为规则的调整和算法的改进提供可靠的数据。白箱评测有助于了解信息处理系统内部组成部分的性能, 又叫做“内在评测”(intrinsic assessment)。

4.2.3 自然语言信息处理系统的语言文字评测基本上只涉及系统的外在的总体性能, 因此, 主要采用黑箱评测的方法。

4.2.4 对自然语言信息处理系统的输入和输出采取有区别的评测态度, 采用“宽进严出”的策略。因为系统可能需要根据用户要求处理不规范的自然语言信息, 系统的输入部分应允许存在不规范之处; 系统的输出部分应严格规范。

5 文语转换系统的评测

语音合成是文语转换的一个部分, 它被包含在文语转换之中, 本标准把语音合成和文语转换系统统称为文语转换系统。

在正式进行文语转换系统评测之前, 一般要对语音编码器进行一些非正式的测试, 主要包括语音编码器的信噪比测试或频谱间距测试等, 这些都是物理性能的测试, 当这些测试都达到要求后, 还要对原始语音进行相互比较, 对各种语音编码器进行相互比较。一般要对语音编码器进行这些非正式的测试之后, 才可以进行正式的文语转换系统评测。

正式的文语转换系统评测包括语音学模块评测和语言学模块评测两部分。语音学模块评测用于评测文语转换系统中的语音合成部分, 整个文语转换评测除了进行语音学模块的评测之外, 还应当进行语言学模块的评测。

语音学模块评测内容包括语音清晰度测试和语音自然度测试。语言学模块评测内容包括切词、多音字、数字串、符号和单位等的文本处理能力的测试。

5.1 语音学模块评测

语音学模块评测内容包括语音清晰度 (articulation) 测试和语音自然度 (naturalness) 测试两部分。语音清晰度是指输出语音是否容易听清楚, 语音自然度是指输出语音听起来是否自然。

语音清晰度测试又可进一步分为音节清晰度测试、单词清晰度测试和单句清晰度测试。

5.1.1 音节清晰度测试

测试方法主要有诊断性押韵测试法 (Diagnostic Rhyme Test, DRT) 和改进的押韵测试法 (Modified Rhyme Test, MRT) 两种。

诊断性押韵测试法: 根据国家标准《GB/T 13504-1992 汉语清晰度诊断押韵测试 (DRT) 法》和《GB/T 16532-1996 通信设备清晰度 DRT 法评价用语音材料库》来进行汉语音节清晰度测试, 测试时, 一般以三个音节为一组, 每个组为一个文本文件。例如, 下面是 DRT 音节表的一部分:

第 1 组是选汉因

第 2 组是英知绑

第3组是狗说更

改进的押韵测试法：MRT 是一种改进的押韵测试法，测试时，一般以一个音节为一组，每个组为一个文本文件。MRT 测试法的结果可以作为 DRT 测试法的参考。例如，下面是 MRT 音节表的一部分：

1 我读元字

2 我读习字

3 我读住字

5.1.2 单词清晰度测试

采用“语义不可预测句” (Semantic Unpredictable Sentence, SUS) 来测试单词清晰度。这样的句子在语义上是不可预测的。如果采用语义可预测句的方式来评测单词清晰度，对听音会有负面的影响，影响测试的准确性。采用语义不可预测的方式，就可以避免根据上文来猜测，并且可以在一定程度上削弱学习效应。

语义不可预测句的格式举例如下：

1	隔壁	电报	懂	扣子
2	少有的	服务	蒸	飞机
3	学问	房子	洗	锤子
4	老太太	拥护	白	道德
5	爷爷	下去	一	碟子

在编制单词清晰度测试题目的时候，应该注意参照《普通话异读词审音表》（1985 年 12 月修订），选择一些《普通话异读词审音表》中规定必须区分读音的单词，例如，在“供给、供销、提供”中的“供”读为[gong1]，在“口供、翻供、上供”中的“供”读为[gong4]；在“银行”中的“行”读为[hang2]，在“行人”中的“行”读为[xing2]。在测试题目的单词中，还应当注意词形的规范，参照《CF1001-2001 第一批异形词整理表》，使用推荐词形。例如，“毕恭毕敬”和“必恭必敬”，应当使用推荐词形“毕恭毕敬”，“奢靡”和“奢糜”，应当使用推荐词形“奢靡”。

5.1.3 单句清晰度测试

测试单句的清晰度，主要测试单句中语调的清晰性和单句类型的差异性，因此，可以采用语义可预测的句子编制句表来进行测试。

在测试单句清晰度时，应该注意区分单句的类型（如，陈述句、命令句、感叹句、疑问句），注意测试标点符号对于句子语音清晰度的影响。

句表的实例如下：

- 1 经济起飞像神话似的创造出了奇迹。
- 2 他成了大学教授，不忘人民养育之恩。
- 3 在人的一生中，机会可遇而不可求！
- 4 通过国旗法进行爱国主义教育。
- 5 实验失败了 3 2 次，难度可想而知。

在编制句表时，应该注意选择一些《普通话异读词审音表》中规定必须区分读音的单词，并注意词形的规范，使用推荐词形。

单句清晰度也可以采用语义不可预测句子来测试，从而避免听音人根据上文猜测下文和学习效应等弊病，以提高单句清晰度测试的准确性。这样的单句清晰度测试句表由若干个语义不可预测的句子组成。这些句子在构成上符合语法，并且尽可能覆盖测试语种的典型语法结构，但在语义上，是独立于上下文的，是不可预测的，因而可以更客观地测试单句的清晰

度。

5.1.4 语音自然度测试

语音自然度测试主要测试合成语音的拟人性、连贯性和韵律感。

- 语音的拟人性：考察合成语音是否与人的语音接近。
- 语音的连贯性：考察合成语音是否连贯，节奏与停顿是否自然，语速是否正常，发音是否流利。
- 语音的韵律感：考察合成语音的语调、重音的位置、轻声和儿化是否正确，听起来是否费力。

这些特征在短文中得到比较全面的体现。因此，语音自然度的测试一般以短文作为测试材料。

短文集合可以根据领域来分类，分为通用领域测试短文集合和特定领域测试短文集合。

5.1.4.1 通用领域测试短文集合：由不同体裁的若干段短文组成，短文长度控制在 50—200 个字以内。

例如，

吃什么，怎么吃？这个本来十分简单的问题，现在正困扰着拥有 5 0 0 0 年历史的中国人。在城市，在沿海地区，在富裕起来的农村，有 8 0 % 的家庭主妇对一日三餐感到头疼。在不断加快的生活节奏冲击中，不能忍受长时间背着做饭重担，一生长在厨房里重复劳动。但是，自己不做饭又到哪里去吃？大的饭店太贵，吃不起；小餐馆的价格倒便宜，可不卫生。

5.1.4.2 特定领域测试短文集合：由若干个欢迎语、提示语、情景对话、信息服务内容的片断组成，它们一般与赛事报道、天气预报、交通、旅游、餐饮服务领域相关。短文长度控制在 50—200 个字以内。

a. 关于交通的短文，例如

从今天开始，北京地铁将延长运营时间，其中 2 号线地铁运营延长时间多达 82 分钟。具体运营时间为：1 号线：苹果园站首班车 5: 10，末班车 23: 30；四惠东站首班车 5: 05，末班车 23: 50。2 号线：积水潭站首班车 5: 10，末班车次日 0: 18。13 号线：西直门站首班车 6: 00，末班车 21: 45。

这段短文中，要测试时间的读法。例如，“23:30”要读为“二十三点三十分[er4 shi2 san1 dian3 san1 shi2 fen1]”，“5: 10”要读为“五点十分[wu3 dian3 shi2 fen1]”（注意：“5: 10”中的“:”为全角符号）。

b. 关于赛事的短文，例如

本主题可以为您提供参赛者、参赛队、体育场馆以及赛程安排的详细信息。您可以通过输入运动员或运动队的名称来获得他们的信息。

在出征巴黎之前，中国队预定的目标是三枚金牌，但在本届世乒赛全部五个单项的角逐中，中国队获得女单、女双、男双、混双四个项目的金银牌，还包揽了混双项目的前四名。男子单打虽然未能获得冠军，但孔令辉、马琳、王励勤都是在关键局的比赛中意外失手，本身具备了获得冠军的实力。

在这些短文中，要测试体育赛事的一些术语的切词是否正确，如“参赛者、赛程、女单、女双、男双、混双”等；人名的切词是否正确，如“孔令辉、马琳、王励勤”等，以保证合成短文的自然度。

c. 关于天气预报的短文，例如

预计，本周全国大部地区以晴到多云天气为主；华北北部和东北地区的气温将比常年同期偏低，我国其它大部地区气温接近常年或偏高。北京地区具体预报如下：9 月 30 日至 10 月 1 日，有降雨，降雨量一般有 3~15 毫米。10 月 2 至 4 日，北京天气以晴为

主，奥运村地区和主场馆温度 19 至 25℃，丰台棒球中心温度 18 至 24℃，昌平赛马场温度 16 至 23℃。

这段短文中，要测试“～”的读法，应该读为“到[dao4]”或者“至[zhi4]”，还要测试温度的读法，例如，“25℃”要读为“二十五摄氏度[er2 shi2 wu3 she4 shi4 du4]”。

5.1.5 语音清晰度评分方法

音节清晰度采用 DRT 或 MRT 方法测试，单词、单句清晰度采用 SUS 方法测试。收集完所有系统的合成语音后，打乱顺序播放，由听音人对合成语音的结果测试进行打分，按百分比进行评分。

a. DRT 和 MRT 测试时，听音人要辨识语音合成结果中的音节，正确辨识的音节数量越大，错误辨识的音节数量越小，音节辨识的效果就越好，因此，最后得分可以按下式计算：

$$P=100\times(R-W)/T$$

其中 R 是正确辨识的音节数量，W 是错误辨识的音节数量，T 是总共测试的音节数量，P 是清晰度辨识的得分。

b. 测试单词清晰度时，听音人要辨识语音合成结果中的单词，应该按单词来判断正误，对于双音节词或多音节词，若其中一个音节错，则判该词为错。

c. 测试单句清晰度时，听音人要辨识语音合成结果中的句子，应该按句子来判断正误，听音人记录每一短句中如有一个关键词错误即判为全句错。

单词和单句清晰度的得分按下式来计算：

$$P=100\times R/T$$

根据这个公式计算得出的 P 的比计算 DRT 和 MRT 测试时得出的 P 要宽一些，因为音节识别比单词识别和单句识别相对容易。

5.1.6 语音自然度评分方法

a. 两两比较评分法 (Paired Comparison, PC)：同一段短文，如果参测系统为 n 个，每个系统的语音合成结果都要和其他的 n-1 系统的合成结果对比，这样共有 $n(n-1)/2$ 个对比组合。m 段短文就有 $n(n-1)m/2$ 个两两对比组合（每个组合由 A 和 B 组成，A, B 表示同一段短文的不同的语音合成系统的合成版本）。听音人按照拟人性、连贯性、韵律感等因素对合成结果采用五级评分制（+2, +1, 0, -1, -2）评分，评分标准如下：

比较结果	A 得分	B 得分
A 比 B 好	+2	-2
A 比 B 稍微好	+1	-1
A 和 B 不相上下	0	0
A 比 B 稍微差	-1	+1
A 比 B 差	-2	+2

b. 平均评价分 (Mean Opinion Score, MOS) 评分法：

MOS 评分法对于听音人要求较高，最好邀请有经验的听音专家来参与评分。听音专家根据语音合成结果的输出与自然语言接近程度的总体印象，从拟人性、连贯性、韵律感等方面，用优、良、中、差、劣五级记分来评价，言语自然度评价的 MOS 分级为：

- 1-劣，不能接受；
- 2-差，不愿接受；
- 3-中，可以接受；
- 4-良，愿意接受；
- 5-优，很自然。

参测系统各段得分平均作为其自然度得分。

评分时，以 PC 法的评分为主，以 MOS 法的评分为辅。

在用 MOS 法评分时，如果听音专家不清楚五级记分的具体表现就可能出现大的方差。可以把好的语音和坏的语音例子让听音专家先听一下，然后再开始测试打分。在同样的测试条件下，在不同国家用本土语言，听音专家不容易在等级定位上取得相互一致。因此，MOS 需要进行调整之后，才可以得到可靠的品质指标。

5.2 语言学模块评测

语言学模块测试是文语转换系统必测的项目。其内容包括：切词、多音字、数字进位制、符号与单位(所有键盘符号、常用化学物理符号及计量单位)等的文本处理能力。具体要求如下：

a. 切词：在经过文语转换系统后输出的拼音文件中，单词边界用空格(space)表示，要特别注意地名和机构名以及术语的切词是否正确，对于汉族人名应当以是否符合韵律词的要求进行判断，姓氏和名字应当连读而不能分读。例如，“毛泽东”应连读为[maozedong]，而不能分读为[mao zedong]。

b. 多音字：参测系统应根据上下文在输出的拼音文件中对多音字给出正确的拼音。例如，“参加”和“参差”中的“参”，前者读为[can1]，后者读为[cen1]。要特别注意单独成词的多音字处理，例如，“为”、“长”等多音字单独成词的时候应当分别根据不同的上下文来读音。“他长于写作”中的“长”读为[chang2]，“他长得很高”中的“长”读为[zhang4]；“拜他为师”中的“为”读为[wei2]，“为人民服务”中的“为”读为[wei4]。

c. 姓氏的特殊读音：参测系统应能区别姓氏的特殊读音。如，“曾国藩”和“曾经”中的“曾”，前者是姓氏，读为[zeng1]，后者读为[ceng2]。

例句：

记者带着这个问题采访了中国食文化研究会会长曾老。这位 75 岁老人曾参加八路军，四面八方都到过。

其中的两个“曾”，第一个“曾”是姓氏，应读为[zeng1]，后一个“曾”应读为[ceng2]。

又如，“仇为之”（人名）和“仇恨”中的“仇”，前者是姓氏，读为[qiou2]，后者读为[chou2]。

例句：

它的地址在旃坛寺，老板姓仇。

其中的“仇”是姓氏，应读为[qiou2]。

d. 数字进位制：对测试材料中的数字串，应按汉语习惯以亿、万、千、百、十为单位读出，如 1,254,000,000 应读成“十二亿五千四百万[shi2 er4 yi4 wu3 qian1 si4 bai2 wan4]”。在输出拼音文本中应给出相应的拼音形式。

例句：

这片林子共有 14,000 棵树。

其中的 14,000 应读为“一万四千[yi1 wan4 si4 qian1]”

e. 年代、时间、电话号码、百分比、分数和小数：参测系统应能区分年代、时间、电话号码和特殊数字表示的顺序式读法和进位制读法以及某些特殊读法，并能处理全角的数字符号。

例句：

食源开发和物种驯化，中国在 4,000 年前就开始进行。

其中的“4,000 年”应读为“四千年[si4 qian1 nian2]”，采用进位制读法。

美联社 16 日报道了中国首位进入太空的宇航员安全返回地面。报道说，在环绕地球 21 个小时后，航天飞船按计划准时着陆。中国的指挥控制中心宣布：中国首次载人航天飞行获得圆满成功。报道说，这次飞行的圆满完成是中国 11 年载人航天计划取得的最高成就，也是中国赢得世界声望的象征。

其中的“16”应读为“十六[shí2 liú4]”，“21”应读为“二十一[èr4 shí2 yī1]”，“11”应读为“十一[shí2 yī1]”，都采用进位制读法。

秦朝建立于公元前 221 年。

其中的“221 年”应读为“两百二十一年[liǎng3 bǎi3 èr4 shí2 yī1 nián2]”，采用进位制读法。

“马克思生于 1818 年。”

其中的“1818 年”应读为“一八一八年[yī1 bā1 yī1 bā1 nián2]”，采用顺序式读法。

“研讨会定于 12 月 23 日上午 9:35 开幕。”

其中的“12”，“23”都采用进位制读法，分别读为“十二[shí2 èr4]”和“二十三[èr4 shí2 sān1]”，“9:35”表示时点，应读为“九点三十五分[jiǔ3 diǎn3 sān1 shí2 wu3 fēn1]”。

“旅游投诉电话是 9258。”

其中的 9258 应读为“九二五八[jiǔ3 èr4 wu3 bā1]”，采用顺序式读法。

有 80% 的家庭主妇对一日三餐感到头疼。

其中的“80%”应读为“百分之八十[bǎi2 fēn1 zhī1 bā1 shí2]”。（注意：80% 是全角的数字符号）

美国太空发展经费占全球约 80.2%。

其中的“80.2%”应读为“百分之八十点二[bǎi2 fēn1 zhī1 bā1 shí2 diǎn3 èr4]”（注意：80.2% 是全角的数字符号）

他的年龄是我的 1/2。

其中的“1/2”应读为“二分之一[èr2 fēn1 zhī1 yī1]”。

2/5 等于 0.4。

其中的“2/5”应读为“五分之二”，“0.4”应读为“零点四[líng2 diǎn3 sì4]”。

我将住 5~8 天。

其中的“5~8”应读为“五到八[wu3 dào4 bā1]”或者“五至八[wu3 zhì4 bā1]”。

f. 符号与单位：对测试材料中的符号和单位，有中文法定计量单位的应给出相应的拼音形式，并按照汉语普通话读音，读音应遵照《关于在我国统一实行法定计量单位的命令》（1984 年）的规定；一般外文符号可按原文给出，按照原文读音。

例句：

1987 年七月肯德基前门餐厅开业，门脸儿招牌上 KFC 三个大字，远远儿就瞧见了。顾客排队最长达 20m，中午就餐最长达 3000-4000 人，真有人驱车 20km 从通县来的，够火的吧！

其中的“20m”应读为“二十米[èr4 shí2 mǐ3]”；“20km”应读为“二十公里[èr4 shí2 gōng1 lǐ3]”。

中国选手获得男子举重 60kg 级冠军。

其中 60kg 的应读为“六十公斤[liú4 shí2 gōng1 jīn1]”

声音在空气中传播的速度是 340 米/秒。

其中的“340 米/秒”应读为“三百四十米每秒[sān1 bǎi2 sì4 shí2 mǐ3 měi3 miǎo3]”。

比热容单位（焦耳每千克开尔文）的国际符号是 J/(kg·K)。

其中的 J/(kg·K) 应按英文字母读音。

g. 以西文字母开头的词语：以西文字母开头的词语有的是借词，有的是外语缩略语，其中的西文字母部分按西文读音，汉字部分按汉语普通话读音。例如，“α 粒子”应读为[ālfā lì4 zǐ0]，“B 超”应读为[B chāo1]，“ATM 机”应读为[ATM jī1]。

h. “一”“不”的读音：现有的用于语音处理的汉语发音词典还没有很好的模型来处理“一”“不”等字的读音。这是因为这些字发音变化的语音上下文环境很复杂。一般在发音

词典中只包含某些最基本的形式（例如“一”的发音为[yi1]），在文语转换时，要使用相应的算法根据上下文推出它们的发音变体。

- “一”在非去声前变为去声
- “一”“不”在去声前变为阳平
- “一”“不”夹在词语中间时变为轻声

h. 上声变调：上声在语流中发生音变，在文语转换时，这种语流音变十分复杂，也要使用相应的算法根据上下文推出它们的发音变体，主要应处理如下的现象。

• 上声在非上声（阴平、阳平、去声）前一律变为平上，调值由原来的[214]变为[21]，只降不升。例如，“影星，影评，影印”中的“影”应读为平上。

• 上声在上声前（上上相连），前一个上声变得像阳平，调值由[214]变为[24]，只升不降。例如，“本领，讲解，导演”中的“本，讲，导”调值为[24]。

i. 轻声的读音：普通话的轻声具有区别意义的作用，在语音合成的评测中，应当注意评测如下要点：

• 辨义轻声：同一个汉字，由于是否读轻声而导致语义不同。例如，“老子”读轻声时表示骄傲的自称，不读轻声时表示古代人名或书名。

• 连接词“和”读为轻声。

• 助词“的、地、得”读为轻声。

• 方位结构中的非中心音节读为轻声：例如，“眼里、手上、乡下”中的“里、上、下”读为轻声。

• 双字重叠的指人名词，后一个音节读为轻声。例如，“哥哥、妈妈、婆婆”中的后一个音节“哥、妈、婆”读为轻声。

• 单音节动词重叠式的后一个音节读为轻声。例如，“看看、洗洗、说说”中的后一个音节“看、洗、说”读为轻声。

j. 儿化的读音：儿化音对于语音合成的自然度有重要的作用，在语音合成中，应当对儿化进行系统化的处理：

• 对于有区别意义作用的儿化词，必须按儿化读音。例如，

信（表示“信件”）—— 信儿（表示“消息”）

头（表示“脑袋”）—— 头儿（表示“领头的人”）

• 对于有区别词性作用的儿化词，必须按儿化读音。例如，

盖（动词）—— 盖儿（名词）

尖（形容词）—— 尖儿（名词）

• 对于表示感情色彩的儿化词，尽量按儿化读音。例如，

小孩 —— 小孩儿

好玩 —— 好玩儿

• 语音词典中，应当对上述儿化词一一标注其拼音，儿化词中的音节数等于汉字字数减一。例如，“花儿”应标注为[hua' er]，其音节数为1。

• 当自动切词得到后缀“儿”时，将“儿”与前面的单词合并，并把前面单词的最后一个音节儿化，语音合成时“儿”不再发音。

• 非儿化词中的“儿”，应当单独读成一个音节。例如，“孤儿、男儿、混血儿”中的“儿”，都应当读成一个音节，不能儿化。

在语音合成评测时，应当考察系统对于上述问题的处理能力。

k. 专有名词的读音：专有名词是文语转换中的一个困难问题；词典中不可能事先列举出汉语中的一切专有名词；专有名词还可能来自其他语言，而且还可能有不同的拼写方法。语音合成和文语转换的很多应用都是与专有名词分不开的；例如，在与电话有关的应用中，

电话簿和打电话都离不开人名和地名。汉语专有名词有的读音很特殊，应该注意区别。例如，“单”作为姓时应读为[shan4]，不能读为[dan1]。地名“枞阳”中的“枞”应读为[zong1]，不能读为[cong1]。

1. 专业术语的读音: 把语音技术应用于不同的专业领域需要正确处理专业术语的读音。例如，地貌学术语“潟湖”（浅水海湾因湾口被淤积的泥沙封闭而形成的湖）中的“潟”应读为[xi4]，不读为[xie4]。

6 语音识别系统的评测

语音识别的目的是让计算机通过识别和理解把语音信号转变为相应的文本或命令。

6.1 测试样本和发音人的选择

进行语音文本转换和音节转换测试时，应当采用包含一定数量句子的汉语普通话样本，由发音人朗读，作为测试数据。测试样本中有标点，在录音时，约定标点不发音，但应有适当停顿。

选择样本的发音人时应当注意：

- a. 适当的性别搭配，以便反映发音时性别的差异：例如，发音人可以男女各半，男 10 人，女 10 人。
- b. 适当反映方言的特点，以便测试计算机是否能够识别带有方言色彩的普通话。

6.2 语音文本转换测试与音节测试

a. 语音文本转换测试的目的是测试非特定人的、无限词汇的、以朗读方式读出的汉语普通话连续语音识别技术，输入是语音信号，输出是相应的文本或者命令；

b. 音节测试的目的是测试语音识别系统的算法。测试时，输入是无调的语音串，输出是相应音节的拼音形式，可以输出多组不同音节来测试算法的功能。

6.3 语音识别的评测指标

6.3.1 语音文本转换评测指标

语音文本转换的评测指标是参照自然语言处理中“词错误率”的概念设计的。

- 汉字正识率 = 正确汉字数/标准答案汉字总数×100%；
- 插入汉字错误率 = 插入错误汉字数目/标准答案汉字总数×100%；
- 删除汉字错误率 = 删除错误汉字数目/标准答案汉字总数×100%；
- 替换汉字错误率 = 替换错误汉字数目/标准答案汉字总数×100%；

例如，对于如下的语音文本转换结果：

标准答案：	历	时	三	天	三	夜	*	顾	不	上	休	息
识别结果：	历	*	三	田	伞	也	勾	顾	布	尚	休	息
评估：		D		S	S	S	I		S	S		

其中，标准答案汉字数为 11，识别结果有 5 个替代（S），1 个插入（I），1 个删除（D），正确识别汉字数为 5：

$$\text{汉字正识率} = \frac{5}{11} \times 100\% = 45.45\%$$

$$\text{插入汉字错误率} = \frac{1}{11} \times 100\% = 9.09\%$$

$$\text{删除汉字错误率} = \frac{1}{11} \times 100\% = 9.09\%$$

$$\text{替换汉字错误率} = \frac{5}{11} \times 100\% = 45.45\%$$

此外，还要评测如下指标：

- 句子正识率 = 正识（一字不差）句子数/句子总数×100%；
- 完成时间 = 系统完成全部任务所用的时间；
- 系统故障数 = 在评测过程中测试系统出现故障的次数。

6.3.2 音节识别评测指标

参照语音文本转换的评测指标，音节识别的评测指标如下：

- 音节正识率 = 正确音节数/标准答案音节总数×100%；
- 插入音节错误率 = 插入错误音节数目/标准答案音节总数×100%；
- 删除音节错误率 = 删除错误音节数目/标准答案音节总数×100%；
- 替换音节错误率 = 替换错误音节数目/标准答案音节总数×100%；

例如，对于如下的音节识别结果：

标准答案：	li2	shi2	san1	tian1	san1	ye4	*	gu4	bu4	shang4	xiu1	xi2
识别结果：	li2	*	san1	tian2	san3	ye3	gou1	gu4	bu4	shang4	xiu1	xi2
评估：		D		S	S	S	I					

其中，标准答案音节数为 11，识别结果有 3 个替代（S），1 个插入（I），1 个删除（D），正确识别音节数为 7：

$$\text{音节正识率} = \frac{7}{11} \times 100\% = 63.63\%$$

$$\text{插入音节错误率} = \frac{1}{11} \times 100\% = 9.09\%$$

$$\text{删除音节错误率} = \frac{1}{11} \times 100\% = 9.09\%$$

$$\text{替换音节错误率} = \frac{3}{11} \times 100\% = 27.27\%$$

此外还要评测如下指标：

- 完成时间 = 系统完成全部任务所用的时间；
- 系统故障数 = 在评测过程中测试系统出现故障的次数。

6.3.3 对语音识别系统产生的文本也须进行语言学方面的评测。其内容包括：字形、异形词、同音词、歧义（切分）、儿化词等的处理能力。具体要求如下：

a. 字形：经过语音识别系统输出的文本在字形方面应符合《第一批异体字整理表》《简化字总表》的规定，使用规范字形。要特别注意计量单位的用字，应使用《部分计量单位名称统一用字表》的规定字形。例如，“十四海里/十四哩/十四海涅”，应选用规范的“十四海里”。

b. 异形词：经过语音识别系统输出的文本对异形词的处理应注意词形规范，使用《第一批异形词整理表》的推荐词形。例如，“百叶窗”和“百页窗”，应选用推荐词形“百叶窗”；“参与”和“参预”，应选用推荐词形“参与”。

c. 同音词：语音识别系统应能对同音词做出正确识别。例如，“枇杷”和“琵琶”读音完全相同，但在“我爱吃枇杷”中，应能识别为“枇杷”；在“她会弹琵琶”中应能识别为“琵琶”。

d. 歧义（切分）：语音识别系统应能对歧义语音串做出正确识别。例如，对“计算鸡蛋

的重量”应能做出正确识别，不能识别为“计算机蛋的重量”。

e. 儿化词：语音识别系统应能对儿化音节进行正确识别，并在文本中给以标识，具体标识方式可以灵活处理，如“把盖儿盖上”、“把盖（儿）盖上”、“把盖_儿盖上”都可以，但如果输出为“把盖盖上”，则认为没有正确识别出儿化音节。