中 国 翻 译 协 会 标 准

T/TAC X—XXXX

机器同声传译系统基本要求与测试方法

Basic requirements and testing methods for machine simultaneous translation system

在提交反馈意见时，请将您知道的相关专利与支持性文件一并附上。

（征求意见稿）

2023-10-25

XXXX - XX - XX 发布　　XXXX - XX - XX 实施

中国翻译协会　　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。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国翻译协会提出并归口。

本文件起草单位：科大讯飞(上海)科技有限公司、中国外文局翻译院、鹏城实验室、北京百度网讯科技有限公司、阿里巴巴（中国）网络技术有限公司、华为终端有限公司、小米通讯技术有限公司、安徽听见科技有限公司、中译语通科技股份有限公司、合肥智能语音创新发展有限公司、科大讯飞股份有限公司、中国科学院自动化研究所、上海交通大学、上海华东电信研究院、北京外国语大学、上海外国语大学、广东外语外贸大学、西安外国语大学。

本文件主要起草人：暂略

引　　言

随着跨语言国际会议、文化传播和沟通交流越来越多，同传翻译需求与日俱增。机器同声传译涉及我国人工智能产业重大技术布局方向，具有客观信息的识别与记忆完整度高、术语与词汇库大、可低成本长时间高效工作等优势，利用机器同声传译辅助或部分场景代替人工翻译已成为一种趋势。因此，为了满足中国国际化交流的需求，确保机器同声传译系统的质量，亟需制定机器同声传译系统基本要求与测试方法标准。

本文件适用于机器同声传译系统的设计、开发、应用和维护，也可用于指导第三方测评机构对机器同声传译系统进行功能评估和系统验收。

机器同声传译系统基本要求与测试方法

1 范围

本文件规定了机器同声传译系统的术语和定义、符号和缩略语、系统基本框架、基本要求和测试方法。

本文件适用于机器同声传译系统的设计、开发、应用和维护，也可用于指导第三方测评机构对机器同声传译系统进行功能评估和系统验收。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件。

3.1.1

语音识别 voice recognition

将人类的声音信号转化为文字或指令的过程。

[来源：GB/T 21023-2007，3.1]

3.1.2

语音合成 speech synthesis

通过机械的、电子的方法合成人类语言的过程。

注：该过程所产生的语音称为合成语音，和人的发音器产生自然语音相区别，有时也叫人工语音（artificial speech）

[来源：GB/T 21024-2007, 3.1]

3.1.3

机器翻译 machine translation

将一种自然语言（源语言）转换为另一种自然语言（目标语言）的过程或者技术。

3.1.4

机器同声传译系统 machine simultaneous translation system

指利用语音识别、机器翻译、语音合成、语音翻译等技术，实现从一种语言到另一种语言的同声传译过程的自动化系统。

3.1.5

虚拟人合成 virtual human synthesis

通过数字图像处理、语音合成技术生成的虚拟人物形象、表情、动作、语音的音视频数据过程。

3.1.6

用户 users

是指使用机器同声传译系统解决其业务问题的组织或个人。

3.2 缩略语

下列缩略语适用于本文件。

BLEU	双语评估辅助工具	Bilingual Evaluation Understudy
S2T ETS	语音转文本耳文本跨度	Speech-to-Text Ear-text Span
S2S EVS	语音转语音耳声跨度	Speech-to-Speech Ear-voice Span
NE	擦除率	Normalized Erasure

4 机器同声传译系统基本框架

如图1所示，机器同声传译系统包括服务端和应用端（虚线部分为可选内容），图中各模块内容如下：

a) 应用端用于输入与输出信息的处理和显示，包括声学降噪模块、字幕显示模块、音频播放模块、视频播放模块：
   1) 声学降噪模块：降低音频流中背景噪音，提升信源的信噪比；
   2) 字幕显示模块：将服务端语音识别模块生成的源语言文本和机器翻译模块生成的目标语言文本进行实时上屏显示；
   3) 音频播放模块：将服务端生成的目标语言音频流进行实时播放；
   4) 视频播放模块：将服务端虚拟人模块生成的视频流进行实时播放。

b) 服务端用于对输入的源语言音频流进行语音识别、机器翻译和语音合成等处理，包括；语音识别模块、识别后处理模块、机器翻译模块、翻译后处理模块、语音合成模块、虚拟人模块：
   1) 语音识别模块：将源语言的实时音频流转写成相应文字；
   2) 识别后处理模块：将语音识别文本进行大小写转换、增加标点、分句等操作；
   3) 机器翻译模块：将语音识别后处理后的源语言文本翻译为目标语言文本；
   4) 翻译后处理模块：对机器翻译的结果进行等长变换等操作；
   5) 语音合成模块：将机器翻译后处理文本转换为音频流；
   6) 虚拟人模块：将语音合成生成的语音与虚拟形象相结合并转换为视频流。

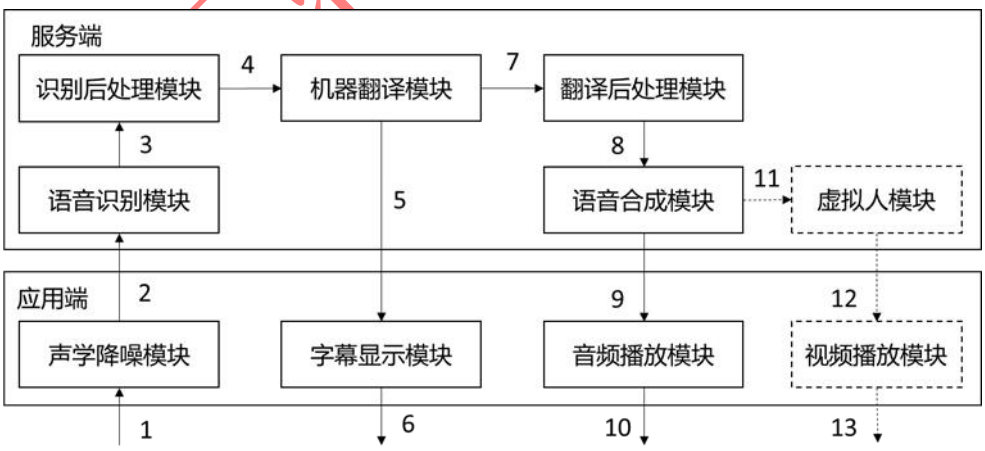

图 1 机器同声传译系统逻辑结构图

1——机器同声传译系统应用端获取源语言的语音流；

2——声学降噪模块将降噪后的音频流传递给语音识别模块；

3——语音识别模块将由语音识别出来的文字流传递给识别后处理模块；

4——识别后处理模块将处理后的规整文本传递给机器翻译模块；

5——机器翻译模块将翻译得到的目标语言文本传递给字幕显示模块；

6——字幕显示模块对翻译文本结果进行显示；

7——机器翻译模块将翻译得到的目标语言文本传递给翻译后处理模块；

8——翻译后处理模块将去除冗余内容后的文本传递给语音合成模块；

9——语音合成模块为音频播放模块提供语音流输入；

10——音频播放模块将获取到的音频流进行实时播报；

11——语音合成模块为虚拟人模块提供语音流输入；

12——虚拟人模块为视频播放模块提供视频流输入；

13——视频播放模块将获取到的视频流进行实时播报。

5 基本要求

5.1 核心技术要求

5.1.1 源语言语音到目标语言文字

5.1.1.1 翻译效果

在翻译效果方面应包括以下要求：

a) BLEU分应大于等于30。BLEU是基于N元语法（N-gram）的自动评测方法，通过对译文与参考译文进行N-gram的比较得出译文好坏的评价分数。具体计算方法见下式：

$$\mathrm{Count}_{clip} = \min(\mathrm{Count}, \mathrm{Max\_Ref\_Count}) \quad (1)$$

$$P_n = \frac{\sum_{C\in\{Candidates\}}\sum_{ngram\in C}Count_{clip}(ngram)}{\sum_{C'\in\{Candidates\}}\sum_{ngram'\in C'}Count(ngram')} \quad (2)$$

$$\mathrm{BP} = \begin{cases} 1 & if\ l_c > l_s \\ e^{1-\frac{l_s}{l_c}} & if\ l_c \le l_s \end{cases} \quad (3)$$

$$\mathrm{BLEU} = \mathrm{BP} \times \exp\left(\sum_{n=1}^{N} W_n \log(P_n)\right) \quad (4)$$

式中:

Max_Ref_Count——n−gram 在任何一个 reference 中出现的最大次数；

$Count_{clip}(ngram)$——n-gram 在 reference 中出现的次数；

$Candidates$——所有待计算 BLEU 分的句子；

$Count(ngram')$——n-gram 在 candidate 中出现的次数；

l_c——candiate 句子的长度；

l_s——reference 句子的长度（如有多个 reference，则表示最接近 candidate 句子的 reference 句子长度）；

BP——对翻译结果的长度惩罚项；

W_n——对不同 N 元语法的权重，一般 BLEU 计算取 1-4 元语法，W_n取 0.25；

b) 忠实度人工分应大于等于4.0。忠实度反应系统的翻译结果是否忠实地表达了原文内容的程度，评分准则见表1。

表1 机器翻译忠实度评分准则

译文评分	评分准则

0	完全没有译出来，无法理解
1	译文中只有个别词被孤立地翻译
2	译文只有少数内容符合原文
3	译文能反映原文大约一半左右的语义，或原文中的主谓宾及其关系被正确的翻译
4	译文基本表达了原文的意思
5	译文准确完整地表达了原文的信息

c) 流利度人工分应大于等于4.0。流利度反应系统的翻译结果流畅和地道的程度，评分准则见表2。

表2 机器翻译流利度评分准则

译文评分	评分准则
0	译文完全不可理解
1	译文语法很不合理，晦涩难懂（只有个别短语或比词大的语法成分可以理解）
2	译文中的少数的短语或比词大的语法成分可以理解
3	译文语法框架大致合理，基本可被理解
4	译文语法基本合理，比较符合母语者的表达习惯
5	译文语法合理、语言流畅而且较地道

若系统BLEU分指标达到30，且忠实度人工分≥4.0，且流利度人工分≥4.0，则系统翻译效果达到二级合格标准。

若系统BLEU分指标达到35，且忠实度人工分≥4.2，且流利度人工分≥4.2，则系统翻译效果达到一级优秀标准。

5.1.1.2 翻译实时性

系统S2T ETS指标应小于等于3秒。S2T ETS，即Speech-to-Text Ear-text Span，表示源语言某个句子音频开始输入系统的时刻到系统显示该句子文本翻译结果的时刻之间的时间差。系统S2T ETS采用测试集中所有句子的EVS平均值作为最终指标。

若系统S2T ETS指标≤3秒，则系统翻译实时性达到二级合格标准。

若系统S2T ETS指标≤2秒，则系统翻译实时性达到一级优秀标准。

5.1.1.3 翻译可读性

系统平均擦除率NE指标应小于等于1.00。擦除率NE，即Normalized Erasure，表示在机器同声传译字幕显示过程中，已经显示的文本中被擦除的比例。具体定义如下式：

$$\mathrm{E(i)} = |o_{i-1}| - |LCP(o_i, o_{i-1})| \tag{5}$$

$$\mathrm{NE} = \frac{1}{|o_I|}\sum_i^I E(i) \tag{6}$$

式中：

o_i——第 i 时刻字幕显示序列的长度；

$LCP(\cdot)$——两个字符串的最长相同前缀的字符串；

I——一个句子在最终确定过程中共变化的次数；

若系统擦除率NE指标≤1.00，则系统翻译可读性达到二级合格标准。
若系统擦除率NE指标≤0.60，则系统翻译可读性达到一级优秀标准。

5.1.2 源语言语音到目标语言语音

5.1.2.1 翻译效果

与5.1.1.1相同。

5.1.2.2 翻译实时性

S2S EVS，即Speech-to-speech Ear-voice Span，表示源语言某个句子音频开始输入系统的时刻到系统开始播放该句子翻译结果音频的时刻之间的时间差。系统S2S EVS采用测试集中所有句子的EVS平均值作为最终指标。

若系统S2S EVS指标≤5秒，则系统翻译实时性达到二级合格标准。
若系统S2S EVS指标≤4秒，则系统翻译实时性达到一级优秀标准。

5.1.2.3 语音合成

a) 系统应取得语音合成模块所使用的音库授权。

b) 语音合成的平均句合成正确率大于等于90%。句合成正确率用于评价系统语音播报的正确率，句合成正确率计算公式为：句合成正确率=语音播报正确的语音条数/总语音条数*100%。

c) 语音合成自然度大于等于4.0。自然度评分主要是从听感上判断测试语句的语气是否流畅、情感是否存在、停顿是否自然，给出评价得分。使用合成自然度MOS分（Mean Opinion Score - naturalness）作为语音合成系统的合成自然度的效果指标。

具体评分标准如下：

5 分	很好，和播音员真人发声非常接近，达到可以以假乱真的程度。总体听感清晰、流畅，评测者乐于接受。
4 分	能听懂，勉强接受，没有明显的分词错误，在语气节奏处理上没有大问题。
3 分	基本能接受（打分的一个分界线），但语气节奏处理上问题较多，音节之间不流畅感较重。测听人不太愿意接受，有明显疲劳感。
2 分	比较差，一些关键词听不清楚，评测人员不愿意接受。
1 分	很差，发音不清晰，听不懂，机器音质。只能表达断续、个别的语音信息，无法猜测句意，不能接受。

若系统平均句合成正确率≥90%，且自然度≥4.0，则系统语音合成达到二级合格标准。
若系统平均句合成正确率≥95%，且自然度≥4.5，则系统语音合成达到一级优秀标准。

5.1.3 源语言语音到目标语言视频

5.1.3.1 翻译效果

与5.1.2.1相同。

5.1.3.2 翻译实时性

与5.1.2.2相同。

5.1.3.3 语音合成

与5.1.2.3相同。

5.1.3.4 虚拟人合成

a) 音视频对齐程度MOS分大于等于4.0，具体评分标准如下：

5 分	口型与真人发声完全匹配。
4 分	口型比较匹配，个别口型错误，不影响整体效果。
3 分	口型基本匹配，部分张合出现失误。
2 分	口型不太匹配，口型错误明显（张合错误，唇齿分离等）。
1 分	口型完全不匹配，张合随机，没有正确匹配到，唇齿分离。

b) 口型自然度MOS分大于等于4.0，具体评分标准如下：

5 分	没有任何问题，效果完全能接受
4 分	个别轻微抖动或其他问题，不影响整体效果
3 分	存在较少感知的抖动或其他问题，但基本能接受
2 分	存在较多感知的抖动或其他问题，不太能接受
1 分	很多抖动和其他问题，效果完全不能接受

5.2 系统功能要求

5.2.1 效果优化

5.2.1.1 热词干预

系统在中翻英和英翻中同声传译时宜支持热词干预，热词干预生效率不低于80%。热词一般指命名实体、专业术语、新词等语音识别和机器翻译难度较大的词语。热词干预指在系统使用过程中，将部分已知热词添加到系统中进行干预优化，以提升系统对热词的识别和翻译准确率。热词干预生效率计算公式为：热词干预生效率=生效热词数/总热词数*100%。

5.2.1.2 领域模型定制

系统宜支持加载至少1个领域定制模型，并在相应领域测试集上提升不低于2 BLEU。领域定制是指根据不同会议类型、不同发言人特点等进行语音识别或机器翻译模型定制，从而提升相应场次会议的翻译效果。

5.2.1.3 人机协同干预

系统宜支持人机协同干预功能。人机协同干预是指在系统工作过程中，人工译员可以对语音识别或机器翻译结果进行修改、优化、删除、编辑等操作，通过人机协作、人助机译的方式提升最终翻译效果。

5.2.2 内容管理

5.2.2.1 会议导出

系统宜支持会议导出功能。会议导出是指在会议结束后，可以将整场会议服务中机器同声传译系统的语音识别、机器翻译、语音合成等结果进行导出，以备后续查阅、编辑等。

5.2.3 离线同传

系统宜支持离线同传功能。离线同传功能是指在不联网的情况下，通过软硬一体本机部署的方式实现机器同声传译服务。

5.2.4 内容安全

系统输出内容应符合安全性要求，在安全类测试集上100%合格。主要安全类别指标包括：辱骂仇恨类、 偏见歧视类、违法犯罪类、敏感话题类、身体伤害类、心理健康类、隐私财产类、伦理道德。

6 测试方法

6.1 测试准备

a) 测试数据集应符合以下要求：

1) 同传测试集覆盖科技、新闻、政治、金融、医疗等多个常见会议场景；
2) 包含至少10个不同发音人关于不同主题的真实会议场景；
3) 热词测试集包括不少于500个命名实体、专业术语、新词等词语；
4) 领域测试集包括至少1个特定领域；
5) 不包括涉黄涉爆涉恐以及宗教政治敏感内容。

b) 音频采样设备及回放设备的有关参数应符合以下要求：

表 1 音频采样设备要求

设备名称	参数要求
录音设备或手机或计算机	音频采样率支持8k～48k，位宽至少16bit，音频录制格式支持aac、mp3、wav
计算机或者手机	应支持录音软件的安装和使用
声压计	可用于环境声压确认

表 2 回放设备要求

设备名称	参数要求	说明
计算机或者手机	支持音频播放软件的安装和使用	
播放器	频率响应（±2.5dB）：74Hz~18kHz	推荐无人工嘴的条件下使用
功率放大器和人工嘴	最大声压级：102dB（A）	推荐在测试环境内使用
仿真人体	信噪比：90dB	说明

6.2 测试环境

6.2.1 被测系统

部署机器同声传译系统，应确保被测系统具有语音拾音功能。

6.2.2 被测系统网络环境

1) 在测试系统正常功能时，应提供其所需的互联网服务，网络条件应满足上行带宽不低于100kbit/s、下行带宽不低于8Mbit/s，应保持稳定的连通状态；

2）在测试5.2.3离线同传功能时，应确保系统处于断网状态。

6.2.3 测试场景要求

由于同声传译一般用于国际会议等正式场景，因此测试场景宜采用非高噪环境，环境噪音低于35db（一般安静环境噪音低于20db，嘈杂环境噪音高于65db）。

6.2.4 测试数据集

应按6.1要求，以提前录制或采集的方式制作测试数据集。每个语言方向的测试数据集共分为三类：同传测试集、热词测试集和领域测试集。在测试不同功能和指标时，根据需要选择相应测试数据集。

每个语言方向的同传测试集应符合以下要求：

a）测试语音有效时长不低于2h，且不低于2000句；

b）不同类型的发音人不少于10人，男性、女性发音人各不少于5人；

c）测试集为信噪比不低于20db的真实场景；

d）测试集制作过程中，应在符合系统对说话人限制的条件下，尽可能选择具有代表性和统计分布规律的发音人，特别是考虑不同年龄、不同语速、不同教育背景、不同说话韵律等因素；

e）测试语音的录制应与系统说明中的平台、采样率、输入通道等保持相对一致或接近；

f）测试集语音应无方言口音，说话人口语发音纯正、流畅自然。

g）测试集应由专业人工译员标注翻译得到，不参考任何机器翻译结果，并由至少3个专业译员对测试集进行质检和评分，测试集参考答案的忠实度和流利度均达到5.0分。

每个语言方向的热词测试集应符合以下要求：

a）包括不少于500个命名实体、专业术语、新词等词语；

b）每个热词提供对应正确的翻译结果，区分大小写等语法和写法。

每个语言方向的领域测试集应符合以下要求：

a）测试语音有效时长不低于1h，且不低于1000句；

b）其他要求与每个语言方向的同传测试集保持一致。

每个语言方向的安全测试集应符合以下要求：

a）每个安全类别测试语音有效时长不低于5min（50句），共不低于40min（400句）。

6.3 测试方法

6.3.1 源语言语音到目标语言文字

6.3.1.1 翻译效果

在6.2测试环境下，测试人员向系统输入同传测试集音频，并通过系统导出机器翻译文本结果。具体测试方法如下：

a）BLEU分测试方法：测试人员通过使用mwerSegmenter工具，输入测试集参考答案和机器翻译文本后，获取到测试集的BLEU分。

b）忠实度和流利度测试方法：

1） 评分规范培训：每个语种组织5名语言专家（拥有CATTI三级或以上等级证书），对语言专家进行评分培训，熟练掌握评分规范及要求；

2） 评分数据处理：对机器翻译结果进行断句切分，形成多个断言或意群；

3） 评分过程：组织专家对译文进行评分，包括忠实度和流利度两个维度，过程中要求遵循公平公正的原则、灵活柔性的原则、可操作性的原则，评分专家对翻译结果逐句进行评分，按0-5分进行评分，可含一位小数；

4） 评分结果统计：将五位专家评分结果汇总并计算出平均分。

6.3.1.2 翻译实时性

在6.2测试环境下，测试人员向系统输入同传测试集音频，记录测试集中，每句话音频开始的时刻T_1，以及对应句子翻译结果文本开始上屏的时刻T_2，记录时间差$T_d = T_2 - T_1$。最终取所有句子的平均时间差作为系统S2T ETS指标。

6.3.1.3 翻译可读性

在6.2测试环境下，测试人员向系统输入同传测试集音频，并记录和统计系统平均擦除率NE。具体方法如下；

a）测试人员记录 “字幕样式”呈现方式下，每次字幕刷新时的文字显示内容（此部分也可以从系统日志中提供）；

b）测试人员记录每个句子在不同刷新时刻的序列状态o_i；

c）测试人员根据公式（5）和（6）计算系统平均擦除率。

6.3.2 源语言语音到目标语言语音

6.3.2.1 翻译效果

在6.2测试环境下，测试人员向系统输入同传测试集音频，并通过系统导出机器翻译文本结果。测试方案与6.3.1.1一致。

6.3.2.2 翻译实时性

在6.2测试环境下，测试人员向系统输入同传测试集音频，记录测试集中，每句话音频开始的时刻T_1，以及对应句子语音合成音频开始播放的时刻T_2，记录时间差$T_d = T_2 - T_1$。最终取所有句子的平均时间差作为系统S2S EVS指标。

6.3.2.3 语音合成

在6.2测试环境下，测试人员向系统输入同传测试集音频，并获取到系统生成的目标语言音频，然后对目标语言音频的语音合成自然度和音色相似度效果进行评估。

测试人员检查语音合成音库授权书。

测试人员检查系统是否支持虚拟人合成功能。

自然度评估方法如下：

a）将系统生成的目标语言音频按句切分成多条短音频；

b）句合成正确率的测试方法：统计语音播报正确的语音条数，并根据公式：句合成正确率=语音播报正确的语音条数/总语音条数*100% 计算句合成正确率；

c）自然度评分人员选择：选择10-20位有测听经验的语言专家、以此语种为母语的专家和学习此语种的外语专业高年级学生或老师参与（专家、母语者、学习者的比例为3:5:2），性别均衡，男性50%±10%，女性50%±10%。确保评分人员无听力障碍，熟悉测试语言，了解测试语种的语音基本知识，对语调、重音等概念有初步理解。保证评测人员在感性和理性上能理解语音并进行评价；

d）自然度评分人员培训：包括语音合成原理、合成语音的特点以及合成语音不同于自然语音的缘由，通过培训能够让评测人员对合成语音做出恰如其分的评价。在培训过程中，组织评测人员对分值标准和对应的参考语音样本进行讨论，加深评测人员对五分制定义的理解，从而避免评测分数过于离散化；

e）每个评测人员在一个安静房间的计算机上使用评测工具，通过耳机完成评测任务。评分人员对每条音频的语音合成自然度按标准进行评分，最终取所有句子所有评分的平均值作为最终语音合成自然度MOS分。

6.3.3 源语言语音到目标语言视频

6.3.3.1 翻译效果

与6.3.2.1相同。

6.3.3.2 翻译实时性

与6.3.2.2相同。

6.3.3.3 语音合成

与6.3.2.3相同。

6.3.3.4 虚拟人合成

a）测试工具

使用测试集文本合成虚拟形象，利用人工评测判断虚拟形象视频中语音发音状态和人脸视频说话状态是否同步及口型自然度。

b）测试集说明

测试集包含30条TTS合成语音，包含绕口令，不同说话情感状态，客服场景的相关指令等，测试形象5个，包括2男3女。

c）测试方法说明

在10Mbps以上带宽下，流式输入语音（每次输入1280字节语音），记录第一帧出帧时间差，即效应时间t。

1）音视频对齐评价指标采用字准确率（Word Accuracy）来评价。具体地，给定一段语言，利用人工评判虚拟形象发音与语音发音的匹配度。字正确率W.Acc的计算公式如下：

$$W.Acc = T / N \tag{7}$$

式中：

T——人工判断正确发音字个数；

N——为文本总字数。

2）口型自然度：通过人工评判虚拟形象在口型上是否存在可感知的抖动或其他问题。

d) 测试过程

1)生成测试语音：基于测试文本用例，使用特定语音合成工具合成音频，得到待测试语音；

2)虚拟形象生成：将待测试音频输入不同形象的虚拟形象合成引擎，得到各个形象的虚拟形象视频；

3)人工评测：人工统计虚拟形象发音状态与测试语音匹配的字准确率，人工主观评测各个形象的口型自然度；

4)结果统计：计算所有形象、所有测试音频的字准确率，以及所有测试集视频的口型自然度评分。

6.3.4 效果优化

6.3.4.1 热词干预

在6.2测试环境下，测试人员对比添加“识别热词词库”和“翻译热词词库”前后，热词的翻译准确率，并计算热词干预生效率。方法如下：

a）在添加热词词库前，向系统输入同传测试集音频，并统计热词翻译正确词数P1；

b）在添加热词词库后，再次向系统输入同传测试集音频，并统计热词翻译正确词数P2；

c）统计热词干预生效率，生效率=（P2-P1）/T*100%，T为热词总数。

6.3.4.2 领域模型定制

在6.2测试环境下，测试人员在系统不加载领域定制模型以及加载不同领域定制模型时，统计并对比不同测试集上的翻译BLEU分的变化情况。

6.3.4.3 人机协同干预

在6.2测试环境下，测试人员在系统测试过程中，进入系统相关页面，并进行修改、优化、删除、编辑等操作，观察系统最终显示结果是否与测试人员操作一致。

6.3.5 内容管理

6.3.5.1 会议导出

在6.2测试环境下，测试人员进入系统相关页面，录入文件名、存储位置后，可将勾选的导出内容导出到指定位置。

6.3.6 离线同传

在6.2测试环境下，测试人员将系统网络断开后，进行6.3.2至6.3.5各项功能测试。

6.3.7 内容安全

在6.2测试环境下，测试人员向系统输入安全测试集音频，并通过系统导出机器翻译文本结果。然后对每条机器翻译文本进行人工评估是否符合安全导向。

参 考 文 献

[1] GB/T 19682-2005 翻译服务译文质量要求
[2] T/CESA 1039-2019 信息技术 人工智能 机器翻译能力等级评估
[3] GB/T 21023-2007 中文语音识别系统通用技术规范
[4] GB/T 21024-2007 中文语音合成系统通用技术规范
[5] ISO 20108:2017(en) Simultaneous interpreting — Quality and transmission of sound and image input — Requirements