计算机科学
COMPUTER SCIENCE

面向算法模型的语音数据集质量评估方法研究

李荪, 曹峰, 刘姿杉

引用本文

李荪, 曹峰, 刘姿杉.面向算法模型的语音数据集质量评估方法研究[J]. 计算机科学, 2022, 49(11A): 210800246-6.

LI Sun, CAO Feng, LIU Zi-shan. Study on Quality Evaluation Method of Speech Datasets for Algorithm Model [J]. Computer Science, 2022, 49(11A): 210800246-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800246

面向算法模型的语音数据集质量评估方法研究

李 荪 曹 峰 刘姿杉
中国信息通信研究院 北京 100191
(lisun@caict.cn.cn)

摘 要 随着智能语音技术和产品应用大规模的成熟落地,对高质量语音数据集的需求与日俱增。目前,针对结构化数据的质量评估方法有一定的研究,但尚未形成面向非结构化的语音数据集质量评估标准。通过研究语音算法模型的构建原理,分析语音数据集的建设需求,建设统一的语音数据集质量评估体系。该评估体系从4个维度对面向算法模型训练的语音数据集进行质量评价,包括广度覆盖性、选集区分性、领域深入性和数据完整性。通过提出具体的语音数据集质量评估指标、计算方法和评估步骤等,对车载应用领域语音数据集的质量进行评估并对结果进行分析,对评估语音数据集质量、促进数据集建设提供参考。考虑了语音数据集构建的多样化适用能力、隐私问题、效率要求、自动化需求等,提出了构建高质量的语音数据集的未来发展建议。

关键词:人工智能;语音数据集;质量评估;算法;模型;智能语音
中图法分类号 TN912.34

Study on Quality Evaluation Method of Speech Datasets for Algorithm Model

LI Sun,CAO Feng and LIU Zi-shan
China Academy of Information and Communications Technology,Beijing 100191,China

Abstract With the maturity of intelligent voice technology and product application,the demand for high-quality voice datasets is increasing. There have been some researchers put effort on the quality evaluation of the structured data,but there are few standards appeared for the unstructured voice dataset. By analyzing the construction principle of speech algorithm model and analyzing the construction demand of voice dataset,a unified quality assessment framework for the voice dataset is presented. The framework proposes to evaluate the dataset in terms of four dimensions,each of which subsumes a set of criteria:breadth coverage,anthology distinction,field depth and accuracy completeness. The criteria that are suitable to evaluate the quality dimensions are presented,each with the definition,measurement method,and the evaluation process for the voice dataset quality measurement. Experimental assessment and analysis results of the voice datasets in the vehicular application field are presented as the reference for evaluating the voice dataset quality,and promoting the construction of the voice dataset. Considering the diversified applicability,privacy issues,efficiency requirements,automation requirements and other aspects of the construction of voice data sets,the development suggestions for building high-quality voice datasets are proposed.

Keywords Artificial intelligence,Speech dataset,Quality assessment,Algorithm,Model,Intelligent speech

1 引言

近年来,人工智能技术快速发展,语音识别的准确率逐步提升,在语音助手、智能音箱、智能可穿戴设备等领域得到了大范围应用。智能语音应用的算法构建和模型训练离不开大量的语音数据集,大规模高质量的语音数据集对语音识别系统的构造具有重要意义。目前,用于人工智能应用的语音数据集质量开始受到越来越多的关注,但已有的大数据质量评估方案大都针对结构化数据,国内外对语音数据集这类非结构化数据集的质量评估,尚未形成系统完整的质量评估体系。

本文首先介绍结构化数据质量、语音数据集质量评价的相关研究成果,通过智能语音识别算法构建对语音数据集的需求分析,指出语音数据集的建设应考虑通用技术性能、应用优化能力与场景适应能力,从广度覆盖性、选集区分性、领域深入性和数据完整性等维度来评价语音数据集的质量。本文提出了体系化的语音数据集评估体系,包括具体维度中适用指标的定义、选取原因、计算方法和评估流程等。选取车载应用领域语音数据集进行数据集质量评估实验与结果分析,对评估语音数据集质量、促进数据集建设提供参考示例。最后,考虑语音数据集构建的多样化适用能力、隐私问题、效率要求、自动化需求等方面,提出了构建高质量的语音数据集的未来发展建议。

2 数据集质量相关研究进展

2.1 数据集质量评价相关研究

随着大数据与人工智能的发展,数据集质量的重要性越来越彰显,并吸引了研究界、产业界与标准化组织投身于相关的研究与数据集的建设当中。

哈佛大学的Meng教授指出,人工智能的数据集质量远比数据量更重要,数据质量评测体系的构建是亟需解决的

通信作者:刘姿杉(Liuzishan@caict.ac.cn)

关键问题。数据集质量评估发展至今，主要还是集中在结构化数据方面，已形成较为丰富的评价模型与标准体系。国际上数据质量整体框架方面的研究最早开始于 20 世纪 90 年代初，麻省理工学院的 Wang 等启动全面数据质量管理计划(Total Data Quality Management，TDQM)[1]，提出了全面数据质量管理方法，包括定义阶段、测量阶段、分析阶段以及改进阶段。随后 Lee 等于 2002 年提出了信息管理质量评价(Assessment Information Management Quality，AIMQ)数据产品质量评估方法论[2]，提供了对数据质量进行评价和差异分析的方法，并最终形成了 TDQM 的体系框架。相关的研究成果对后来数据质量领域的发展产生了深远影响。该团队还提出了指导数据质量指标定义的一般性原则的设计保证(Design Quality Assurance，DQA)方法[3]。该方法定义了主观和客观评估、主观和客观评估对比以及改进 3 个阶段。在该方法中，数据质量指标大多被定义为特定的，即用于解决特定问题的指标，因此数据质量指标取决于所考虑的问题。我国于 2008 年成立了信息质量研究组(Information Quality Research Group，IQRG)。此外，“数据质量管理的基础理论与关键技术(编号：F020204)”被列为国家自然科学基金委信息科学部 2011 年的年度重点研究项目。北京大学的唐世渭教授和他的研究小组使用六元组的形式来描述数据质量评估模型[4]，提出了构造模型和计算指标的方法。文献[5]提出了基于元数据驱动的数据质量评估体系架构。总的来说，国内数据质量研究仍以局部分散为主，缺乏系统性成果和规模性组织，缺乏针对我国语音数据特点的面向具体领域的数据集质量研究。

发展至今，数据质量相关标准包括 ISO/IEC 25012、ISO/IEC 25024、GB/T 25000.12-2017 和 GB/T 25000.24-2017 等。其中，GB/T 25000.24-2017 所提出的数据质量评估指标主要包括 15 大类，例如数据完整性、有效性等，每个大类下又进行了相关指标的细分。然而，相关标准规范主要针对结构化数据集，难以满足非结构化语音数据集质量评测的需求。

2.2 语音识别模型的训练原理及数据构建研究

语音识别指，以语音数据集为训练数据，主要对语音信号进行分析，将其转换为文字序列的过程，如图 1 所示。语音识别算法模型[6]，首先通过信号处理模块提取解码器需要的特征向量，将声音转化为计算机可以识别的数字序列或向量，接着根据提取的特征序列在解码器中寻找最优解。其中，解码器中声学模型是将语音信号的观测特征与句子的语音建模单元联系起来，通过训练大量的语音数据得出每一帧和状态所对应的概率，语言模型结合概率输出计算出概率最大的文字序列。发音字典则包含系统所能处理的单词的集合，为声学模型的建模单元和语言模型建模单元间的映射关系，组成了一个搜索的状态空间用于解码器的解码工作。最后，在声学模型、语言模型、发音词典共同组成的网络中解码出得分最高的序列，即认为是识别出来的结果。可见，在模型训练和应用过程中，需要考虑模型声学环境、语言内容、语言知识等多方面要素，只有充分考虑通用技术性能、应用优化能力与场景适应能力，才能构建出适用于算法模型的语音数据集。

在传统语言学领域，针对语料库的设计和构建已经出现大量的基础研究工作[7]，这为语音数据集构建和评价提供了充分的依据。在语料库挑选中，需以尽可能少的语料覆盖尽可能多的语音现象，语料挑选的基本算法是贪婪算法[8]。虽然有部分针对中文语音语料库构建和评价方面的研究，但其目的主要是为了语言知识分析和语言应用发展，与算法模型对语音数据集的需求不一致，无法直接应用于数据集质量的评价。

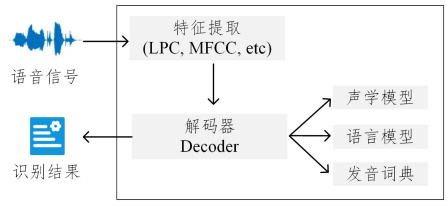

图 1 语音识别技术框架

Fig. 1 Speech recognition technology framework

3 数据集质量相关研究进展

基于语音模型训练原理和语料库构建相关需求分析，针对语音数据本身的语音单元及语音知识，结合模型应用场景，本文提出了面向算法模型训练的语音数据集质量评价体系，包括广度覆盖性、选集区分性、领域深入性和数据完整性 4 个维度，如图 2 所示。

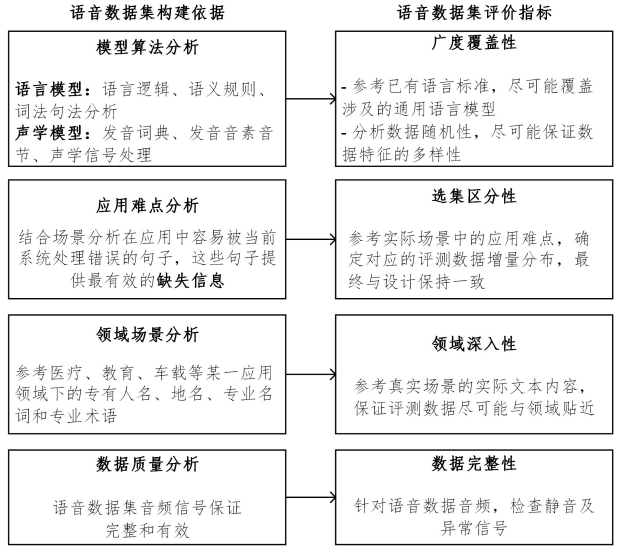

图 2 面向算法模型的语音数据集质量评价体系

Fig. 2 Quality evaluation system of speech dataset for algorithm model

3.1 广度覆盖性

一个大规模语音识别系统需要面对各种各样可能的识别场景，这些识别场景如果在训练语料中没有出现，则识别器在该场景中的识别效果将急剧下降。这意味着，为了构造一个好的识别系统，其训练数据中必须覆盖实际应用场景中的所有可能的环境和条件。因此，有两个方面的广度覆盖性主要评价指标和方法，包括对语言现象的覆盖和数据特征的覆盖。

3.1.1 语言覆盖度

语音数据集应通过积累各种发音条件、发音方式与语法结构，来尽可能地反映真实语言现象。以汉语为例，在发音方面，汉语轻声以《普通话水平测试用必读轻声词语表》为准，共计 137 个；汉语儿化词以《普通话水平测试用儿化词语表》为准，共计 176 个。在结构方面，语言中的词性包括了名词、动词、形容词、数词、量词和代词等各类实词，以及副词、介词、

连词、助词、拟声词和叹词等各类虚词；句法又可以分为主谓结构、偏正结构（修饰、限制关系）、述宾结构（关涉关系/述宾关系）、补充结构（补充关系）、并列结构（联合关系/并列关系）等。

语言覆盖度，主要评估语音数据集中涉及到语言学相关基础信息的覆盖程度，其评估方法为：

$$X=\frac{A\cap B}{B} \tag{1}$$

其中，A 代表指定数据信息样本出现的数目，B 为参考数据样本的总数目；$X\in[0,1]$，X 值越大，说明指定数据样本的覆盖率越大。

3.1.2　特征信息量

语音数据集最重要的是积累各种数据源，让模型可以满足各种变动性，反映产品应用的鲁棒性，须覆盖实际应用场景中所有可能的环境、各种发音条件与发音方式，包括但不限于：1）说话人；2）录音设备；3）传输信道；4）环境噪声；5）场景氛围；6）方式与情绪；7）地域口音；8）专业与领域等。以上每种条件都涵盖更为复杂的子条件，如地域口音可能包括轻微口音普通话、重口音普通话、地方方言、外国人汉语发音等。因此，语音数据集建设的需求之一是积累多样化数据源，从而让模型能够具备足够的泛化能力，反映出算法模型识别的鲁棒能力。

信息熵可以作为特征信息量来度量数据集中特定语音特征的复杂程度。当语音特征越复杂，特征出现不同类别的情况越多样化，信息熵越大；如果语音特征越单一，出现不同种类的情况越少（极端情况只有 1 种类别），此时的信息熵越小（对应上述情况的信息熵为 0）。语音数据集本身存在多维度、多层次、多关联性和重复性等许多复杂特征，最大熵值为每条语音数据特征多样性的体现[9-10]。因此，将每个类型样本数据的信息熵值和最大熵的比例作为评估语音数据集的特征信息量，对于任一特征 X，其计算式为：

$$I=-\sum_{i=1}^{n}\frac{p(x_i)\log_2 p(x_i)}{Y} \tag{2}$$

其中，数据样本总数目为 n，$p(x_i)$ 代表特征 X 取元素值为 x_i 的概率，Y 代表最大熵值；$I\in[0,\log_2(n)]$，I 的值越高，代表特征 X 所提供的信息熵越高。

3.2　选集区分性

在累积语音语料数据源时，需要考虑具体应用场景的真实情况，对能够为算法模型带来特征增量价值的数据进行积累。例如，在语音识别场景，需考虑包含多种噪声和多人对话等复杂场景下的语音数据；在语音合成应用中，需积累中英文混杂的语音数据；在声纹识别场景中，语音数据集需要包含合成语音、转换语音和录音，从而提高算法模型的抗攻击能力等。

为了提高模型的识别率与鲁棒性，在进行语音数据积累时，应加强对当前系统识别容易出错的数据的标注与积累。这意味着选择数据不仅考虑数据本身的价值，更需要考虑数据对系统带来的增量价值，与应用场景的特征分布增量保持一致。根据待评估语音数据集特征分布与所述应用场景特征分布，本文提出利用分布一致性评价方法来进行特征匹配度的评估。针对特征类型为数值类型的语音数据集，计算所述语音数据集的特征与所述应用场景要求的特征的欧氏距离，如语音时长、采样率单数值特征，具体如下：

$$S(Data_1,Data_2)=\sum_{i=1}^{n}\omega_i\times(Data_1{}^{(i)}-Data_2{}^{(i)})^2 \tag{3}$$

其中，假设数据集的特征维度为 N，$[1,n]$ 表示特征类型为数值类型的特征。$Data_1^{(i)}$ 表示待评估数据集的第 i 个特征，$Data_2^{(i)}$ 表示实际应用场景数据的第 i 个特征，ω_i 为对应特征的权重。

针对特征类型为分布类型的语音数据集，分布一致性评价则是通过计算所述语音数据集的特征与所述应用场景要求的特征的 KL 散度来实现，例如频率分布、性别分布、年龄分布等，具体如下：

$$\mathrm{E}(Data_1,Data_2)=\sum_{i=n+1}^{N}\omega_i\times\Big(\sum_{j=1}^{M^{(i)}}Data_1{}^{(i)}(x_j)\times\log\Big(\frac{Data_1{}^{(i)}(x_j)}{Data_2{}^{(i)}(x_j)}\Big)\Big) \tag{4}$$

其中，$[n+1,N]$ 表示数据集中特征类型为分布型的特征，$Data_1^{(i)}$ 表示待评估数据集的第 i 个特征，$Data_2^{(i)}$ 表示应用场景数据的第 i 个特征，ω_i 为对应特征的权重。在分布特征中，每一个特征由分布来表示，假设其有 $M^{(i)}$ 个维度，每一个维度为 x_j。最终，将分布距离平均值作为待评估语音数据集与应用领域数据集的特征匹配度。

3.3　领域深入性

《国家语委现代汉语语料库》构建时参照行业属性，按照领域内容进行划分，包括人文与社会科学类（划分为 8 个大类和 30 个小类）、自然科学类（含农业、工业、医学、电子、工程技术等）和综合类（应用文和难于归类的语料）[11]。类比人工智能模型训练学习，广泛积累多样化语音数据源只能提高人工智能模型的通用识别能力，但对于应用领域所处的专业背景，如法律、电信、医疗、家居、金融等相关应用模型的适应能力，还需要对特定领域的问候语、情景对话、人名、地名、专业术语等专业内容进行语音语料数据源的积累。

可见，语音数据和领域内容的贴合程度成为了领域深入的重要衡量指标，本文采用 TFIDF-COS 算法，对文本进行词频抽象分解，从数据角度量化相似性[12]。评价语音数据集领域深入性，采用内容相似度来衡量，具体方法如下。

首先，计算待评估语音数据集的词频向量，以及预设领域数据集的词频向量[13]。计算语音数据集 x 的词频向量 $\boldsymbol{u}$，计算式如下：

$$\boldsymbol{u}=\frac{f_n}{f}\cdot\log\frac{N}{p_n} \tag{5}$$

其中，f_n 为某词在数据集中出现的次数，f 为数据集中最高频词出现的次数，N 为通用语料库文章的篇数，p_n 为包含该词的文档数。例如，搜索主流搜索引擎，Google 发现，包含“的”字的网页共有 250 亿张，假定这就是中文网页总数。包含“中国”的网页共有 62.3 亿张，包含“蜜蜂”的网页为 0.484 亿张，包含“养殖”的网页为 0.973 亿张。

然后，基于余弦相似度算法，计算待评估语音数据集和预设领域数据集的特征词频向量的相似度[14]。假定 P 和 Q 是两个 n 维向量的数据集，P 与 Q 的余弦相似度计算式为：

$$N=\frac{\sum_{i=1}^{n}(p_i\times Q_i)}{\sqrt{\sum_{i=1}^{n}(p_i)^2}\times\sqrt{\sum_{i=1}^{n}(Q_i)^2}} \tag{6}$$

其中，$\boldsymbol{P}$ 表示待评估数据集的词频向量；Q 表示预设领域数据集的词频向量；$N\in[0,1]$，N 的值越大，说明两个词频向量越接近，文本相似度越高。

3.4 数据完整性

语音数据集在采集的过程中，由于设备、环境和说话人的原因，会造成信号缺失、录音不完整等问题，会产生很多无效信息。语音数据集通常为连续的采样点，信号异常指信号受到采音设备或者环境噪声的影响，出现不连续的语音信号片段；静音片段通常为说话人和录音设备的问题，语音数据集会出现静默信号片段，这些都为语音数据集的无效片段。

对于语音信号质量检测，通常基于语音边界检测(Voice Activity Detection，VAD)方法，可获取所述语音数据集中的信号无效片段[15]。评价语音数据集的完整性，采用内容有效度来衡量，具体表达式如下：

$$X=1-A/B \tag{7}$$

其中，A 代表语音数据集中无效数据时长；B 为待评估语音数据集时长；$X\in[0,1]$，X 值越大，说明数据内容的有效度越高。

4 语音数据集质量评估实验与分析

为了给出利用本文所提出的语音数据集质量评估指标进行质量评估的过程，本文选择公开的具体语音数据集进行质量评估与分析。

4.1 实验步骤

依据以上研究，语音数据集质量评估主要分为 3 个步骤，如图 3 所示，具体如下：

(1)准备待评估语音数据集，包括语音音频、转写文本和元数据特征说明(采集设备、应用领域、发音人等)，以及应用领域语音数据集采集标准和参照文本。

(2)按照语音数据集质量评价模型，分别计算语言覆盖度、特征信息量、特征匹配度、内容相似度和信号有效度。

(3)针对不同的应用类型和训练目标，对语音数据集质量的需求侧重与总体评价目标有所不同。因此，质量评估指标的筛选应该对评价目标有足够的覆盖面，同时与评价目标保持高度的一致性。当所有已选择的评估度量计算完成之后，待评估数据集在每个评估度量上都形成了量化指标评分。

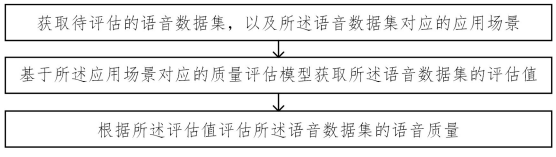

图 3 语音数据集质量评估流程

Fig. 3 Voice dataset quality assessment process

4.2 实验准备

本文选取开源语音数据集 AISHEL1(OPENSLR-SLR33)和 THCHS-30(OPENSLR-SLR18)进行信息统计，结合语音语料库构建要求[16]，设计车载应用领域语音数据集采集标准和文本，具体设定包括：1)内容设定，即车载驾驶，涉及车载器件、驾驶操作、地点名称等内容，包含驾驶模式选择、故障预警、电话服务、语音导航、交通路况播报等汽车驾驶时的语音服务场景；2)语速设定，偏慢(3 秒钟 9 字以下)30%，中等(3 秒钟 9—13 字)50%，偏快(3 秒钟 13 字以上)20%；3)噪声设定，低噪(55 分贝以下)占比 70%，中噪(55～60 分贝)占比 20%，高噪(61～70 分贝)占比 10%；4)发音人设定，男女比例 1∶1，口音标准普通话二级乙以上。

4.3 实验结果

4.3.1 语言覆盖度评估

统计《现代汉语词典第七版》《普通话水平测试用必读轻声词语表》《普通话水平测试用儿化词语表》中声母、带调韵母、轻声、儿化、音节、双音子和三音子的种类数 A，以及待评估语音数据集各类语言现象种类数 B，按照计算公式得到每个语音现象的覆盖度，求均值得到总体语言覆盖度，如表 1 所列。从结果可以观察到，在语言现象的覆盖程度上来说，AISHEL1 数据集的平均覆盖度为 0.731，THCHS-30 数据集的平均覆盖度为 0.573 2，AISHEL1 相比 THCHS-30 包含更丰富的语言发音要素，更能反映实际中文语言发音现象。

4.3.2 特征信息量评估

由于开源数据集中只有发音人的特征信息，因此将发音人个体作为特征，计算不同发音人发音条数占比率的信息熵值与最大熵比例。其中，AISHEL1 数据集覆盖发音人共 400 人，输出音频数 141 908 条，分布如图 4 所示。对于 400 个发音人，根据信息熵计算公式得出所能提供的最大信息熵值为 17.117，AISHEL1 的发音人信息熵为 2.601，对应的特征信息量为 0.151 9。THCHS-30 数据集覆盖发音人共 60 人，输出音频数 13 388 条，分布如图 5 所示。对于 60 个发音人，根据信息熵计算公式得出所能提供的最大信息熵值为 3.710，THCHS-30 的发音人信息熵为 1.743，对应的特征信息量为 0.127 1。在语音特征信息量上，AISHEL1 数据集相比 THCHS-30 数据集提供了更多的发音人语音信息特征，数据源更加多样化，为模型训练提供了丰富的泛化能力和鲁棒能力。

表 1 语言覆盖度统计值

Table1 Language coverage statistics

	AISHEL1			THCHS-30		
	A	B	cov1/%	A	B	cov1/%
声母	21	21	100.00	21	21	100.00
带调韵母	185	168	99.40	185	168	98.81
轻声	154	137	80.29	110	137	52.55
儿化	1	176	0	1	176	0
音节	1 355	1 859	69.77	1 208	1 859	61.00
双音子	15 057	9 853	86.38	8 254	9 853	62.71
三音子	187 696	67 281	75.87	35 613	67 281	26.16
语言覆盖度	—	—	73.10	—	—	57.30

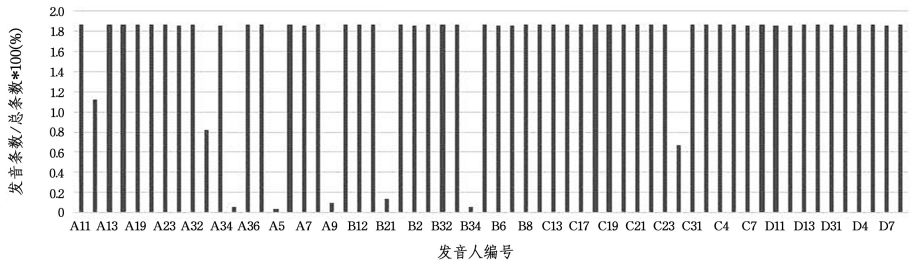

图 4 AISHEL1 发音人发音条数分布统计

Fig. 4 Distribution statistics of number of pronouncements in AISHEL1

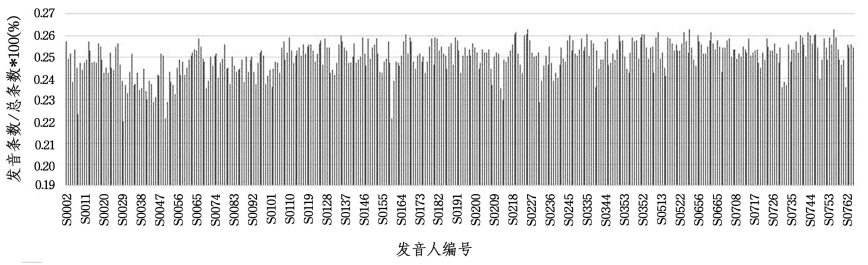

图 5　THCHS-30 发音人发音条数分布统计

Fig. 5　Distribution statistics of number of pronouncements in THCHS-30

4.3.3　特征匹配度评估

依据实验准备中的采集标准，通过规定语速、噪声、发音人等多类要素形成标准型参考分布。但是，AISHEL1 和 THCHS-30 开源数据集并未对以上特征做出标注，无法直接统计信息。但是，信噪比、语速可通过语音活动检测（Voice Activity Detection，VAD）算法来进行统计计算。

经过计算，AISHEL1 数据集的信噪比分布特征匹配度 D1＝0.243 064 205，语速特征匹配度 D2＝0.154 901 96，均值为 0.198 9；THCHS-30 数据集的信噪比分布特征匹配度 D1＝0.536 289 422，语速特征匹配度 D2＝0.154 901 96，均值为 0.345 6。在特征分布匹配程度方面，THCHS-30 数据集相比 AISHEL1 数据集更符合理论语音模型数据构建标准。

4.3.4　内容相似度

分析构建车载领域语音数据集参照文本（3 062 句）的词频及文本内容，如图 6 所示。

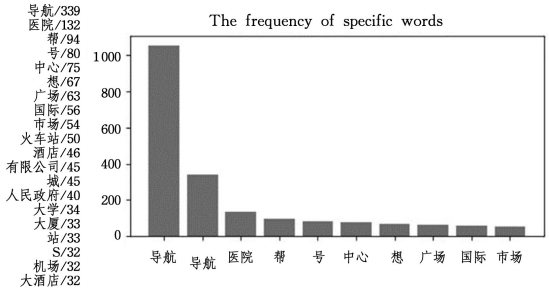

图 6　车载领域语音数据集参照文本分析

Fig. 6　Reference text analysis of voice data set in vehicle field

采用 TFIDF-COS 算法对数据进行全样本关键词抽取和词频分析，结果如图 7 和图 8 所示。经过计算，AISHEL1 数据集的内容相似度为 3.15%，THCHS-30 数据集的内容相似度为 2.18%。可见，在车载领域，THCHS-30 数据集和 AISHEL1 数据集的领域相关度都较低。

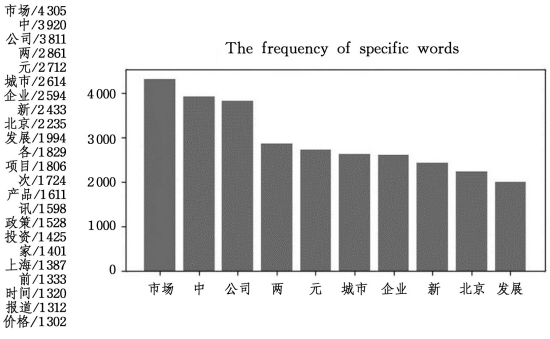

图 7　AISHEL1 语音数据集词频分析

Fig. 7　Word frequency analysis of AISHEL1

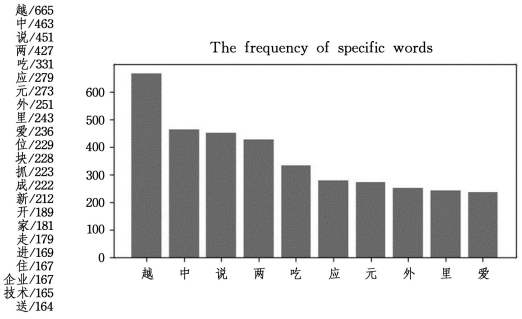

图 8　THCHS-30 语音数据集词频分析

Fig. 8　Word frequency analysis of THCHS-30

4.3.5　信号有效度

利用 VDA 算法来统计去除静音段和异常语音的语音数据集的有效数据量，以评估内容有效性。其中，AISHEL1 数据集语音时长比均值为 70.5%，标准差为 6.4%。其中，24 条存在截幅，截幅比例为 0.017%，除去静音段和异常信号，内容有效度为 0.704 9，时长比分布如图 9 所示。THCHS-30 数据集语音时长比均值为 71.8%，标准差为 5.6%。其中，3 138 存在截幅，截幅比例为 23.4%，除去静音段和异常信号，内容有效度为 0.484，时长比分布如图 10 所示。从语音信号内容有效性方面进行分析，可以得出 AISHEL1 数据集相比 THCHS-30 数据集信号完整度更高。

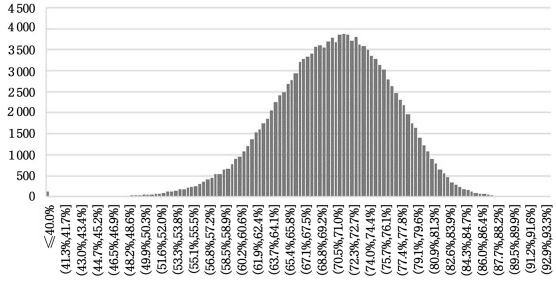

图 9　AISHEL1 语音数据时长比分布

Fig. 9　Voice data duration ratio distribution of AISHEL1

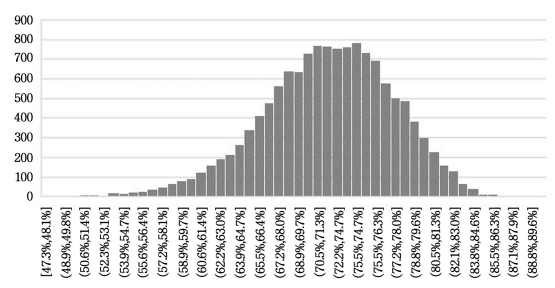

图 10　THCHS-30 语音数据时长比分布

Fig. 10　Voice data duration ratio distribution of THCHS-30

4.4 实验总结

从以上实验结果可以看出，开源语音数据 AISHEL1 和 THCHS-30 在语言覆盖度、特征信息量、特征匹配度、内容相似度和信号有效度几个指标维度上的表现不尽相同，按照平均值计算 AISHEL1 质量相对高一些，如表 2 所列。

表 2 数据集质量评估结果统计

Table 2 Statistics of data set quality evaluation results

指标名称	AISHEL1	THCHS-30
语言覆盖度	0.731 00	0.573 20
特征信息量	0.151 90	0.127 10
特征匹配度	0.198 90	0.345 60
内容相似度	0.031 50	0.021 80
信号有效度	0.704 90	0.484 00
平均值	0.363 64	0.310 34

但是，此次实验还缺少对各个指标权重配比方面的研究，还需要进一步做深入的理论分析和实验。

结束语 本文提出了统一的语音数据集评估方法，包括数据集质量的评估维度与定义、数据集质量评估指标和评估方法等，为评估语音数据集质量、促进数据集建设提供了可靠依据。然而，在实际的应用中，为了构建高质量的语音数据集，还存在一些亟需解决的问题，包括语音数据集构建的多样化适用能力、隐私问题、效率要求、自动化需求等。基于以上分析，针对高质量语音数据集的构建，本文提出了以下发展建议。

(1)建立完善的语音数据集质量评估体系：本文针对智能语音算法模型所需语音数据集，提出了多项可参考的评估指标和方法。未来还需要将理论指标进行泛化，建立完善的评估指标和标准、测评准则与方法，以满足不同应用场景对语音数据集质量的要求，为高质量语音数据集的设计与构建提供重要支撑与参考，这是保证语音数据集质量必不可少的技术要素。

(2)设计自适应的语音数据集质量评估框架：针对不同的智能语音应用场景，同一语音数据集对模型的训练性能表现差异较大。因此，如何设计一个通用、高效并具备自适应能力的语音数据集质量评估框架，构建面向应用的数据集支撑“底座”，是高质量语音数据构建与发展必不可少的组成部分。

(3)在多目标均衡下实现高质量语音数据集的构建与发展：一方面，高质量语音数据集的获取与开放受到了数据安全和隐私保护的制约，数据安全与隐私保护技术往往会影响到数据集的质量，因此在实际应用中需要考虑数据集质量与隐私保护之间的均衡；另一方面，大规模高质量语音数据集的构建在人员、环境、工具等方面需求较高，因此在实际应用中需要考虑数据集带来的信息增量与模型性能优化与成本之间的均衡。

参 考 文 献

[1] WANG R Y, STOREY V C, FIRTH C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640.

[2] LEE Y W, STRONG D M, KAHN B K, et al. AIMQ: a methodology for information quality assessment[J]. Information & Management, 2002, 40(2): 133-146.

[3] PIPINO L L, LEE Y W, WANG R Y. Data quality assessment[J]. Communications of the ACM, 2002, 45(4): 211-218.

[4] YANG Q Y, ZHAO P Y, YANG D Q, et al. Research on data quality evaluation method[J]. Computer Engineering and Application, 2004, 40(9): 3-4, 15.

[5] HUANG G, YUAN M, WU X Y, et al. Research on metadata driven data quality evaluation architecture[J]. Computer Engineering and Application, 2013(8): 114-119, 181.

[6] SHAN Y H, LI J, WANG X R, et al. The generation method of speech recognition training data and the training method of speech recognition model: CN111402865A[P]. 2020.

[7] ZU Y Q. Corpus design of Chinese continuous speech database[J]. Journal of Acoustics, 1999(3): 236-247.

[8] WU H, XU B, HUANG T Y. Automatic corpus selection algorithm based on triphone model[J]. Journal of Software, 2000, 11(2): 271-276.

[9] ZHUANG J L. Research and application of quantitative analysis of data quality[D]. Shanghai: Donghua University, 2019.

[10] JIN J. Research on the value evaluation of information entropy in the era of big data[D]. Changchun: Jilin University, 2019.

[11] LIU L Y. Development of modern Chinese Corpus[J]. Language Application, 1996(3): 3-9.

[12] GHEITH M, ABOUL-ELA M, ARAFA W. Learning Word Graph Representation for Document Classification[C]// 27th Conference for Computer Science, Statistics and Operation Research. Egyptian Computer Society, 2002.

[13] GOU H W, GOU X T. Analysis of word separation and sentence similarity based on word vector[J]. Scientific and Technological Innovation, 2018(33): 55-56.

[14] GU B, LI J H, LIU K Y. Chinese text clustering based on COSA algorithm[J]. Chinese Journal of Information, 2007, 21(6): 65-70.

[15] LIU P, WANG Z Y. Multimodal speech endpoint detection[J]. Journal of Tsinghua University (Natural Science Edition), 2005(7): 896-899.

[16] WANG T Q, LI A J. Design of continuous Chinese Speech Recognition Corpus[C]// National Conference on Modern Phonetics. 2003.

LI Sun, born in 1988, postgraduate, engineer. Her main research interests include machine learning, perceptual cognitive technology and data governance, etc.

LIU Zi-shan, born in 1992, Ph. D. Her main research interests include network intelligence, federated learning, data security and privacy, etc.