ASR Bundestag：大型德语政治辩论数据集

Johannes Wirth 和 René Peinl

霍夫应用科学大学，地址：Alfons-Goppel-Platz 1，95028 霍夫，德　国

摘要。我们介绍 ASR Bundestag，这是一个用于德语自动语音识别的数据集，包含 610 小时经过对齐的音频—转录文本对，可用于监督训练；另有 1,038 小时未标注音频片段，可用于自监督学习。这些数据基于德国议会全体会议和委员会会议的原始音频数据及转录文本。此外，我们讨论了自动创建语音数据集所采用的方法，并通过评估和对预训练最先进模型进行微调来评估所得数据集的质量。我们公开提供该数据集及其所有子集。

关键词：自动语音识别、数据集、德语。

1　引言

随着当前神经网络的参数数量不断增长，训练稳健且性能良好的模型需要越来越多的数据 [1]。尤其是在自动语音识别（ASR）领域，当前的模型架构（如 Wav2Vec 2.0 [2, p. 2] ）基于大量未标注数据进行自监督训练。然而，通常仍需进行微调，且采用监督方式，这需要大量有标注数据；在涉及领域专用词汇或自发语句的应用场景中尤其如此，因为这类语句往往语法不规范。Wirth 和 Peinl [3] 的分析表明，使用更多有标注数据训练的模型（如 Conformer-CTC 和 Citrinet）优于 Wav2vec 2.0；后者是在大型未标注数据集上训练，并仅使用少量有标注数据进行微调的。HuBERT X-Large 目前是规模最大的自动语音识别模型之一，拥有十亿个参数 [4] ，但似乎仍未充分训练，这促使其 [5] 将训练数据进一步扩展至 9.6 万小时，尽管其大型模型仅有 3.16 亿个参数。Chan 等人 [6] 通过 SpeechStew 策略表明，将来自不同领域的大量转录语音合并用于监督训练，可以得到泛化能力良好的模型；该模型在多个数据集上的词错误率（WER）接近针对特定领域的最先进水平（SOTA）模型。英语语音识别数据集数量极为丰富，涵盖朗读语音和自发语音，例如 SPGISpeech [7] 有 5,000 小时转录音频，Gigaspeech [8] 甚至有 10,000 小时；相比之下，德语数据集则相对较少，这一点从训练所得的这些语言的最先进水平模型的性能对比中也可见一斑。

我们推出 ASR Bundestag，这是一个用于语音识别的数据集，由德国联邦议院的政治辩论组成，并包含用于德语语音识别模型监督训练和无监督训练的子集。该数据集涵盖大量自发语音以及政治术语和短语，有助于提升模型在德语上的性能。

本文结构如下：首先概述目前可免费获取的德语语音数据集，随后介绍本文所述数据集的生成过程。之后将更详细地介绍该数据集的特性和子集，并展示利用该数据集进行微调实验的结果。最后，我们将对这些结果进行解读，并在结论中总结前文内容，同时展望未来。

2 用于语音识别的德语数据集

由于德语被视为高资源语言，因此有多个可用于语音识别的数据集。其中一些专为自动语音识别（ASR）设计，例如 M- AILABS [9] 或 Mozilla Common Voice（MCV）[10], ；另一些则来自语音合成项目，例如 HUI [11] 和 Thorsten Voice neutral [12]。还有一些数据集是为语言学研究而创建的，例如 BAS ALC[13] 和 SC10 [14] ，但仍可用于自动语音识别训练。

此外，现有数据集可分为朗读语音和自发语音。后者还可进一步分为对话和演讲。演讲可视为朗读与自发语音之间的连续切换，因为演讲内容可能经过一定程度的编写，但通常不会直接照稿朗读，而且至少包含一部分自发内容。朗读语音的示例包括 Voxforge [15] 和口语维基百科语料库 [16]。自发语音数据集包括德语 TED 演讲 [17] 和 VoxPopuli [18]。然而，对话数据集仍然稀缺，数据量也有限。25 小时的 Hempel 数据集 [19] 以及近期由 Datatang 为 Interspeech 发布的 20 小时子集 2022 [20] 是少数值得关注的例外。会话语音面临的一项挑战，是如何处理口语表达和吞音音节。将这类语音映射为标准德语文本（例如 [20] 所采用的做法），实际上是把两个步骤混为一谈：先将音频进行语音转写，转换为字符；再将方言或口语转换为标准德语文本。使用这类数据进行训练，可以提升自动语音识别模型在这一特定数据集上的性能，但同时也可能降低其在大量数据集上的整体性能。理想情况下，这类数据集应提供两种独立的转写：一种是标准德语文本，另一种是语音准确转写。

有些数据集也关注录制语音的技术方面，例如 Tuda-De [21] 和 Verbmobil 2 [22]。它们分别提供使用不同麦克风同时录制的语音，以及通过手机录制的语音，以便针对这方面对自动语音识别系统进行专门训练或评估。

规模最大、包含超过 3,000 小时德语语音的数据集是 MLS [23], ，由 Facebook 提供。与 M-AILABS 和 Voxforge 一样，它基于 LibriVox [24]。

然而，由于缺少对德语中数字和缩写的规范化处理，其质量相当低。[3]。

最后，可以通过 SL100 [25] 和 Hempel [19]等数据集测试性别偏见，这些数据集包含数量相等的女性和男性说话者。用于训练的大型数据集（尤其是 MCV）并非如此，因此有报告指出存在性别偏见，使女性相较于男性处于不利地位 [26]。令人惊讶的是，Garnerin 等人在评估中发现，即使女性在训练数据中并未代表不足，这种性别偏见依然存在。

表 1。德语自动语音识别和文本转语音数据集与德国联邦议院数据集的比较。

数据集	描述	小时数
BAS ALC	162 名说话人，清醒和醉酒状态	95
Hempel	3920 名说话人，来自不同州，47% 为男性，自发语音	25
BAS SI100	101 名说话人，50% 为男性，朗读报纸文本	32
SC10	70 名说话人，朗读语音和非提示语音	12
VerbMobil 2	自动语音识别，多语言，401 位说话人，会话语音	22
德语 TED 演讲	自动语音识别，德语，71 位说话人，自发语音	16
MCV 7.0	自动语音识别，多语言，15,620 位说话人，朗读语音	1,062
TUDA-De	自动语音识别、德语、147 位说话人、朗读语音	127
SWC	自动语音识别、多语言、363 位说话人、朗读语音	285
M-AILabs	自动语音识别、多语言，基于 LibriVox 等资源	237
MLS	自动语音识别，多语言，244 位说话者，基于 LibriVox	3,287
Voxforge	自动语音识别，多语言，基于 LibriVox	35
HUI	文本转语音，德语，122 位说话者，基于 LibriVox	326
Thorsten	文本转语音，德语，1 位说话者，朗读语音，非常清晰	23
VoxPopuli	自动语音识别，多语言，来自欧盟议会的政治辩论	268
Bundestag Clean		610
Bundestag 脏数据	自动语音识别，德语，来自德国联邦议院的政治辩论	766
Bundestag Unlabeled		1,038

除了 MLS 之外，德语中没有类似 SPGISpeech [7]（含 5,000 小时）或 Gigaspeech [8]（含 10,000 小时转录音频）这样的大规模 ASR数据集。即便是 MLS [23]，其德语数据质量也存疑，而且英语数据远多于德语数据（71,506 小时，德语仅 3,287 小时）。因此，最好的模型在 LibriSpeech 测试干净集及其他数据集 [27], 上的词错误率（WER）分别达到 1.4% 和 2.5%，并不令人意外；而在 M-AILABS 和 Tuda-De 上的词错误率仍然相对较高，分别为 4.3% 和 5.8%。这两个数据集是德语自动语音识别领域引用最多的数据集之一 [3]。

本文通过提供一个名为 ASR Bundestag 的高质量语音识别新数据集，旨在推动德语自动语音识别的发展。

3 数据获取与处理

为创建该数据集，我们首先收集了德国联邦议院会议的录音和转录文本，将其处理为较短的音频片段和规范化的转录文本，随后使用两种不同的处理方法将二者对齐，生成短音频片段与转录文本对。

3.1 数据获取

德国联邦议院的媒体资料库 [28] 向公众提供全体会议和委员会会议的录音，并允许将其用于非商业用途 [29]。该数据集以德国联邦议院的 331 场已录制并转录的会议为基础，这些会议包含 1,200 小时的全体会议录音以及委员会会议录音，并按照下文所述的方法进行了处理。

3.2 数据预处理

由于所有音频均以 44.1 kHz 采样率录制为立体声，并以 MP3 格式存储，因此将采样率转换为 16 kHz、声道数从立体声转换为单声道，并将文件格式转换为 WAV，以确保音频数据可用作所考虑的语音识别模型的输入。

此外，使用内部开发的规范化工具对转录文本的大小写、缩写和数字进行了规范化。在 331 场可用会议中，有 320 场用于生成数据集，其余 11 场因缺少转录文本而被排除。

3.3 对齐方法

在后续步骤中，采用两种不同的方法对音频与转录文本片段进行对齐，进一步处理此前积累的音频转录文本对。这些音频录音时长很长，平均为 3.9 小时。

3.3.1 基于时间戳和 CTC 的切分与对齐

由于所有使用的录音的转录文本均包含时间戳及其他元数据（如说话人变更、掌声等），因此首先利用这些信息切分音频录音，再通过 NVIDIA NeMo 工具包提供的 Citrinet 模型进行转录。该模型在 MCV7.0、MLS 和 VoxPopuli 上进行了预训练， [30] 用于强制对齐转录文本片段与转录文本预测。将原始德国联邦议院转录文本与自动语音识别模型的转录文本预测进行比较后发现，标注的转录文本与实际发言内容相差很大。这在一定程度上是因为重复词语或句子结构经过了修正，但主要原因是转录文本中的时间戳准确度差异显著。

因此，基于时间戳的切分方法加入了时间容差（即在每段音频的开头和结尾各额外增加五秒

片段），以提高转录文本 完全包含在按时间容差切分的音频录音中的可能性。随后，利用采用CTC解码方案的模型进行转录文本预测，并匹配CTC输出，以再次缩小先前设定的时间容差。在此过程中，确定自动语音识别模型的贪心输出标记与相应转录文本开头相匹配的时间步，并移除音频录音中此前的时间步。类似地，采用相同技术，移除转录文本中最后一个标记所在时间步之后的多余音频。

虽然所述强制对齐方法能够在给定的转录文本与音频片段之间生成精确对齐，但它并不适用于创建 ASR数据集，因为转录文本与实际说出的内容差异过大，无法实现语音与转录文本的精确对齐。这主要是由于自发语音中包含口误、口吃和不正确的句子结构，而这些内容在转录文本中会被删除或修正。这些情况使转录文本与音频录音的对齐变得复杂。

3.3.2 基于说话人分离和静音区间的切分与对齐

为了尽可能精确地获得与音频录音中实际说出的发言和句子相匹配的音频转录文本，首先使用用于说话人分离的神经网络 [31](pyannote.audio 2.1.1) 的输出预测结果，对较长的会议录音进行切分；对于这一分离与切分过程生成的、时长超过 30 秒的分段输出，则根据静音阈值再次进行切分。切分操作在 +60 dB 至 -60 dB 的范围内生成，并保留平均时长最长的切分结果以供后续处理。随后，使用预训练的语音识别模型（如 3.3.1 节所述的 Citrinet）为得到的音频片段生成转录文本预测，并将其用作对齐的参考文本。

在强制对齐过程中，会将每个转录文本预测与转录文本中的一部分进行匹配，使字符错误率（CER）达到最低。具体做法是遍历整个转录文本，确定一段文本，其长度与转录文本预测相同，并在其前后各额外包含五个词，从而使字符错误率最低。随后，使用 3.3.1 中所述的基于CTC的收窄来确定转录文本的起始和结束位置；但在本例中，拟合的是转录文本候选项，而不是音频记录。

我们本可以使用[32] 中所述的 CTC 分段来代替这里介绍的匹配过程，但我们发现，所使用的自动语音识别模型存在持续的转录错误，尤其是在处理[3], 中的自发语音时。这些错误表现为“幻觉”：某些词出现在转录文本预测的开头，却并未在转录音频中说出。这是使用持续含有错误的数据进行预训练所导致的症状。这些幻觉可能会在对齐过程中造成严重惩罚。此外，我们选择不采用这种方法，因为德国联邦议院的数据主要包含自发语音，而 CTC 分段最适合处理朗读文本。

4 德国联邦议院数据集

下一节将详细介绍通过前文所述的对齐方法生成的数据集的元数据，以及根据其他指标筛选出的子集。

4.1 基于时间戳和 CTC 的对齐

由于此前提到的局限性，使用时间戳和 CTC 对齐生成的数据集无法直接用于德语语音识别模型的高效训练，也未使用该数据集进行预训练或微调实验。尽管如此，该数据集仍向公众开放，因为它提供了语境准确的音频转录文本对，可用于其他潜在应用。

4.2 基于静音区间和说话人分离的对齐

基于最初的 1,200 小时原始数据，我们创建了一个数据集，其中包含 530,419 组音频转录文本对，总时长为 1,038 小时。由于对齐过程旨在尽量降低转录文本片段与转录文本预测之间的 CER，因此音频转录文本对中也包含质量欠佳的对齐结果；这些结果通过设定最大偏差阈值并采用进一步的算法予以过滤。实验和抽样结果表明，当转录文本与参考文本（转录文本预测）之间的最大 CER 不超过 15% 时，通常可获得高质量的对齐结果。

然而，转录文本片段的开头和结尾处常常会出现相关音频片段中并未说出的词。在检查 Voxpopuli 数据集 [18]，时也发现了类似问题；该数据集是通过类似的基于阈值的筛选（使用 20% 字符错误率）创建的。为尽量减少这种影响，我们为音频转录文本片段设定了另外两个条件，以将其判定为对齐良好：转录文本与转录文本预测的首词和末词相同，或者两者包含相同数量的词。

上述对音频转录文本片段的筛选，将整个数据集划分为对齐质量各异的几部分，以下分别称为“干净”“脏数据”和“未标注”。

干净：
如前文所述，如果能够找到一个与转录文本参考的字符错误率偏差小于 15% 的转录文本片段，并且转录文本与转录文本参考包含相同数量的词，或首词和末词完全匹配，则将音频转录文本对归入干净类别。此外，少于五个词的转录文本会被排除，因为在转录文本相对偏差相同的情况下，较长的录音具有更高的置信度，而且非常短的音频转录文本片段不应包含在该子集中。这一排除也反映在该子集的音频时长分布中（图 1）；相比之下，超集脏数据（图 2）和未标注（图 3）的分布没有设置此类限制。使用经此方法处理的数据进行的初步微调实验取得了成功，但发现共有 ~2% 的离群值，随后对这些离群值进行了调查、排除，并将其纳入

超集“未标注”中。此外，对首词和末词未完全匹配的对齐进行了评估，并进行了半自动校正。符合此前设定限制的音频转录文本对占总数据的 58.8%（~610 小时）。数据集被随机划分为训练集、验证集和测试集（90/5/5），并将平均音频样本时长相近（± 0.01 秒）以及样本时长分布相近作为附加条件。

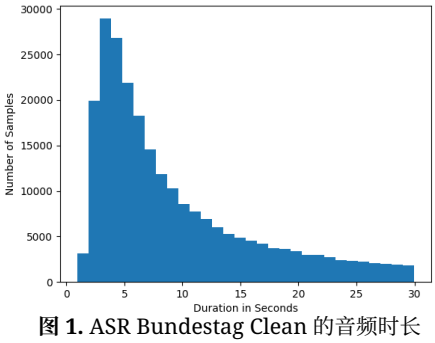

图 1. ASR Bundestag Clean 的音频时长

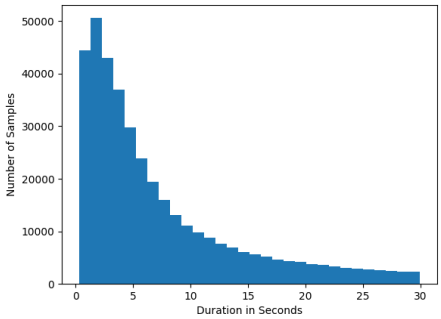

图 2. ASR Bundestag Dirty 的音频

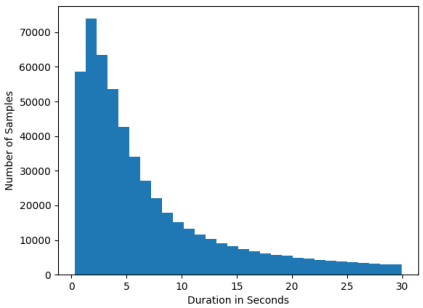

图 3. ASR Bundestag Unlabeled 的

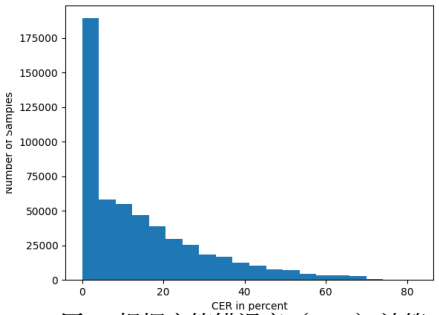

图 4. 根据字符错误率（CER）计算

脏数据：对于此

子集，转录文本与转录文本预测之间的偏差程度提高至 20% CER，且不再对转录文本长度设置下限。该子集仍有条件地适用于自动语音识别模型的训练数据集。为便于今后比较两个数据集，CER 阈值根据 Voxpopuli（转录语音）处理部分所述的筛选流程设定。放宽限制后的这部分构成了“干净”子集以及总数据中额外子集的超集。该集合中的音频转录文本对占总数据的 ~73.8%（~766 小时）。

未标注：
这个涵盖所有内容的超集包含脏数据，因此也包含干净子集，被称为“未标注”，因为它还包括所有自动生成的音频-转录文本

配对，即使其中一些无法确定转录文本与参考文本的对齐，且字符错误率低于20%（完整分布见图4）。在该数据集中，音频片段中所说内容与相应转录文本高度匹配的平均概率过低，无法用它进行监督训练来训练出稳健的自动语音识别模型。不过，它可用于半监督和无监督训练。此集合中所有音频片段的总时长为 1,038 小时。

5 语音识别结果

我们使用 ASR Bundestag 干净版对 Citrinet 模型进行了微调。该模型此前曾用于数据集创建过程中对齐阶段的转录文本预测；此外，我们还微调了一个基于相同训练数据集预训练的 Conformer-CTC 模型。预训练所用的基本配置保持不变，仅修改了以下参数：

- 调度器仍使用余弦退火，仅将最大学习率从 $5x10^{-2}$ 降至 $1x10^{-4}$。
- 移除了预热。
- 为适配训练数据集，将最大输入音频时长从 20 秒增加到 30 秒。
- 批次大小设为 8，以适配所用的图形处理器。
- 训练时长固定为 30 个训练轮次（Conformer-CTC 为 50 个训练轮次）。

这些流程使用 NVIDIA RTX A6000 图形处理器执行。

表2：在 ASR Bundestag 干净版数据集上微调的自动语音识别模型的性能结果和元数据。

模型架构	Citrinet	Conformer-CTC
预训练数据集	MCV7.0、MLS、VoxPopuli	
训练轮次数量	30	
微调时长	87h	76h
预训练词错误率平均值 %	10.68	13.14
预训练词错误率中位数 %	10.00	11.90
微调词错误率均值 %	8.48	9.41
微调词错误率中位数 %	7.14	8.00
绝对提升 %	2.20	3.73
相对提升 %	20.60	28.39

如表2所示，微调使两个模型的词错误率相对降低了20%以上。不过，中位数词错误率明显低于平均值，这可能表明少数转录文本预测的词错误率极高，但总体而言，模型微调后的表现优于平均值所显示的水平。

两个经过微调的模型在德国联邦议院数据集的测试数据上，词错误率均有显著的相对提升，尽管两者此前都已使用相当大量的数据进行预训练。这些提升表明，德国联邦议院数据集中的音频与转录文本片段之间始终保持着准确的对齐。

由于训练结束前损失值仍在持续收敛，因此进一步训练很可能会带来更大的提升。不过，在开展更多微调实验之前，应先进行超参数优化。

深入的错误评估显示，微调纠正了此前转录错误的单词中的 ~30%。此外还发现，年份仍被转录为整数，而不是年份的读法（“one thousand nine hundred ninety” < > “nineteen hundred ninety”）。这是因为模型预训练所使用的 Voxpopuli 数据集始终未能正确规范化年份。虽然 ASR Bundestag 转录文本中的年份已得到正确规范化，但微调过程尚不足以在这一程度上纠正模型。

6 讨论

对数据集创建过程的描述表明，创建高质量的 ASR 数据集需要付出大量努力。必须采用更精细的过滤机制，才能区分那些与实际话语非常接近的参考转录文本，以及对实际所说内容进行较自由改写的其他文本。尽管通过人工质量保证步骤发现潜在问题后进行了多轮迭代和改进，但由于干净数据集上的词错误率中位数与平均值差异显著，所提供的转录文本与音频对齐质量仍存在一些疑问。未来研究需要对此作进一步分析。不过，微调后词错误率显著降低，表明该数据集的整体质量仍然很高。

对齐过程中设定的字符错误率上限可能会限制其在进一步训练自动语音识别模型中的使用，因为现有模型最难处理的复杂术语可能会被排除在 ASR Bundestag 数据集的干净子集之外。然而，若不进行大规模且因此不具备合理性的人工质量保证过程，我们没有发现其他能够确保数据集高质量的可行方法。更粗略的自动对齐过程或许能够纳入更多具有挑战性的词汇，但会显著降低音频与转录文本片段之间精确匹配的置信度。

此外，使用 ASR Bundestag Clean 进行预训练任务的影响尚未得到研究。由于 ASR Bundestag 中音频样本的最大长度比用于预训练微调模型的数据高出 50%，因此，在训练尚未初始化的模型架构时，损失收敛可能会受到负面影响。

此外，目前尚未与其他测试数据集进行交叉评估。尤其值得比较德国联邦议院自动语音识别与 VoxPopuli，因为二者在词汇、录音质量和自发语音方面具有非常相似的特征。

7 相关工作

除第二节引用的数据集论文外，还有许多其他研究也涉及自动语音识别。语音识别的新架构大多与英语数据集一同提出并进行评估。最近，成功应用于大语言模型的混合专家架构也被用于自动语音识别评估。SpeechMoE [33] 在包括 AIShell 在内的四个不同数据集上，相较于计算量匹配的稠密基线模型，错误率相对降低了 7% 至 23%。SpeechMoE 2 [34] 通过改进路由进一步提高了模型准确率；在不同汉语口音的评估中，这一改进对四川方言尤其有效。

在其他几篇文章中，研究者提出了现有架构的更高效版本。Tiny transducer [35] 致力于在边缘设备上使用小型转导器模型进行语音识别。该模型仅有 1.6 M 个参数，在一个具有挑战性的车载噪声数据集上，词错误率从 21.5% 降至 18.1%，比混合时延神经网络基线绝对提升了 3% 以上。Q-ASR 模型 [37] 使用 QuartzNet 和 Jasper 的量化版本，以 8 位而非 32 位表示；尽管其大小仅为基线的四分之一，性能仍几乎与基线相当。

SpeechBrain [38] 和 NeMo [39] 等工具包也致力于提升语音识别结果的准确率。SpeechBrain 在论文中展示了基于现有架构取得的多项新的最先进性能，尤其是在 TIMIT 数据集以及法语和意大利语的 Common Voice 上。NeMo 使用了一个基于 LibriVox 的自有俄语数据集，其中包含 98 小时的语音。除 MCV 外，使用该数据集还使他们能够将测试数据上的准确率从 12.6% 提升至 9.7%，优于仅使用 MCV 的基线。

Bermuth 等人在 Scribosermo 项目中采用了类似的方法 [40]。他们在 37 个德语数据集上训练 QuartzNet 模型，以便在当时的德语 Common Voice 子集上达到最先进水平的词错误率。

Sinha 和 Siegert 改进了德语语音助手对话的语音识别 [42]。他们使用 Voice Assistant Conversation in the wild (VACW) 和 Voice Assistant Conversation Corpus (VACC) 数据集，并利用 ROVER 算法将三种不同云端自动语音识别引擎（Google、IBM、Wit.ai）生成的转录结果合并为一份高质量转录文本。对于 VACC，他们将词错误率从表现最佳的单个自动语音识别引擎的 4% 降至使用 ROVER 算法时的 3.3%。对于 VACW 数据集，他们未能在词错误率方面取得改进。

Lopez 等人最近对荷兰语、英语和德语的会话语音进行了评估 [43]。他们发现，会话语音给语音识别带来了多项挑战，包括“um”和“hm”这样的填充词、弱化现象（在德语中尤为常见，例如用“’n”代替“ein”，或用“ham”代替“haben”），以及截短词，例如用“nich”代替“nicht”。

8 结论与展望

本文介绍了 ASR Bundestag，这是一个用于德语语音识别的大型数据集，包含大量说话人说出的政治术语和短语，并根据音频与转录文本对齐质量划分为多个子集，可用于监督训练和自监督训练。干净子集时长达 610 小时，是规模最大的德语语音识别标注数据集之一，也是规模最大的专注于政治辩论的数据集。此外，我们还公开了基于两种不同分段流程生成的全部音频转录文本对，可通过 https://opendata.iisys.de/ 获取。我们使用预训练的最先进德语语音识别模型架构提供了基线评估，并给出了微调实验的结果。这些实验的积极结果表明，所述对齐方法具有很高的精度，且数据集质量优良，因此，扩充高质量德语有标注语音数据的目标已实现。

未来的工作将通过交叉评估和对更多架构进行预训练，进一步评估数据集的质量。我们还将审视 ASR Bundestag 与 VoxPopuli 之间的质量差异。此外，计划系统研究使用 ASR Bundestag 数据集对自动语音识别模型进行微调的影响，并评估其在其他数据集上的表现，因为增加训练量可能会提升整体性能，也可能导致模型更加专门化，从而降低其在其他数据集上的性能。

参考文本

[1] J. Kaplan 等，《神经语言模型的缩放定律》，*arXiv*预印本 *arXiv:2001.08361*，2020年。[2] A. Baevski、H. Zhou、A. Mohamed 和 M. Auli，《wav2vec 2.0：语音表征自监督学习框架》，*arXiv:200611477*，2020。。[3] J. Wirth 和 R. Peinl，《德语自动语音识别：详细错误分析》，发表于*2022*年*IEEE*全层智能系统国际会议（*COINS*），2022，页码1–8。[4] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《Hubert：通过掩码预测隐藏单元进行自监督语音表征学习》，*IEEE/ACM*音频、语音与语言处理汇刊，第29卷，页码3451–3460，2021年。[5] S. Chen 等，《Wavlm：面向全栈语音处理的大规模自监督预训练》，*arXiv*预印本 *arXiv:2110.13900*，2021年。[6] W. Chan、D. Park、C. Lee、Y. Zhang、Q. Le 和 M. Norouzi，《Speechstew：简单混合所有可用语音识别数据来训练一个大型神经网络》，*arXiv*预印本 *arXiv:2104.02133*，2021年。

[7] P. K. O’Neill 等，《SPGISpeech：用于全格式端到端语音识别的5,000小时转录金融音频》，arXiv预印本 arXiv:2104.02014，2021年。[8] G. Chen 等，《Gigaspeech：包含10,000小时转录音频、不断发展的多领域自动语音识别语料库》，arXiv预印本 arXiv:2106.06909，2021年。[9] I. Solak，《M-AILABS语音数据集》，2019年。https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/ [10] R. Ardila 等，《Common Voice：超大规模多语言语音语料库》，发表于《第十二届语言资源与评测会议论文集》（页码4218-4222），2020年3月。[11] P. Puchtler、J. Wirth 和 R. Peinl，《HUI德语音频语料库：高质量文本转语音数据集》，发表于第44届德国人工智能大会（KI2021），柏林，德国，2021年。[12] T. Müller，《Thorsten开放式德语语音数据集》，2021年3月25日。https://github.com/thorstenMueller/deep-learning-german-tts（访问日期：2021年3月26日）。[13] 巴伐利亚语音信号档案馆，BAS酒精语言语料库。http://hdl.handle.net/11022/1009-0000-0001-88E5-3。2016年。[14] V. Mapelli，Strange语料库10 - SC10（“口音II”）（ELRA-S0114），欧洲语言资源协会（经由CLARIN虚拟语言观测站），国际语言资源编号024-991-750-952-3。2004年。[15] K. MacLean，Voxforge。[在线]。地址：http://www.voxforge.org [16] T. Baumann、A. Köhn 和 F. Hennig，《口语维基百科语料库集合：采集、对齐及其在超听中的应用》，《语言资源与评测》，第53卷，第2期，页码303–329，2019年。[17] F. Hernandez、V. Nguyen、S. Ghannay、N. Tomashenko 和 Y. Esteve，《TED-LIUM 3：数据量翻倍及用于说话人自适应实验的语料库划分》，发表于国际语音与计算会议，2018年，页码198–208。[18] C. Wang 等，《VoxPopuli：用于表征学习、半监督学习和解释的大规模多语言语音语料库》，发表于第11届自然语言处理国际联合会议，2021年，页码993–1003。[19] C. Draxler 和 F. Schiel，《巴伐利亚语音信号档案馆的三个新语料库——迈向分布式网络录音的第一步》，2002年。[20] P. Pan，《应对如何识别和理解高质量人工智能数据解决方案将如何提升模型性能》，发表于Interspeech 2022，韩国仁川，2022年。[在线]。地址：https://www.interspeech2022.org/program/industrytalk.php [21] S. Radeck-Arneth 等，《开源德语远场语音识别：语料库与声学模型》，发表于国际文本、语音与对话会议，2015年，页码480–488。[22] 德国联邦教育与研究部和德国航空航天中心项目管理机构，VERBMOBIL II - VM CD21.1 - VM21.1（ELRA-S0034-30）。欧洲语言资源协会（ELRA），1.0，国际语言资源编号837-421-490-699-3。2004年。[23] V. Pratap、Q. Xu、A. Sriram、G. Synnaeve 和 R. Collobert，《MLS：面向语音研究的大规模多语言数据集》，Interspeech 2020，页码2757–2761，2020年。[24] 《LibriVox | 公有领域免费有声读物》。https://librivox.org/（访问日期：2022年12月15日）。

[25] 巴伐利亚语音信号档案馆，《BAS Sl100》，1995年。http://hdl.handle.net/11022/1009-0000-0007-E9CF-A（访问于2022年3月28日）。[26] M. Garnerin、S. Rossato 和 L. Besacier，《探讨自动语音识别训练数据中性别代表性的影响：以 LibriSpeech 为例》，载于第三届自然语言处理性别偏见研讨会，2021年，第86–92页。[27] Y.-A. Chung 等，《W2v-bert：结合对比学习和掩码语言建模进行自监督语音预训练》，*arXiv* 预印本 *arXiv:2108.06209*，2021年。[28] 《德国联邦议院一媒体库》。https://www.bundestag.de/mediathek（访问于2022年12月15日）。[29] 《议会电视音视频资料使用条款》。https://www.bundestag.de/mediathek（访问于2022年12月15日）。[30] O. Kuchaiev 等，《NeMo：使用神经模块构建人工智能应用的工具包》。arXiv，9月 13, 2019 1909 09577 。[31] http://arxiv.org/abs/。H. Bredin 等，《pyannote.audio：用于说话人分离的神经网络构建模块》。arXiv，11月 04, 2019 1911 01255 。[32] http://arxiv.org/abs/。L. Kürzinger、D. Winkelbauer、L. Li、T. Watzel 和 G. Rigoll，《用于德语端到端语音识别的大型语料库 CTC 分段》，*arXiv:2007.09127*，2020年，doi：10.1007/978-3-030-60276-5_27。[33] Z. You、S. Feng、D. Su 和 D. Yu，《Speechmoe：通过动态路由专家混合扩展至大型声学模型》，*arXiv* 预印本 *arXiv:2105 03036*，2021年。.[34] Z. You、S. Feng、D. Su 和 D. Yu，《Speech moe2：改进路由的专家混合模型》，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2022年，第7217–7221页。[35] Y. Zhang、S. Sun 和 L. Ma，《Tiny transducer：一种适用于边缘设备的高效语音识别模型》，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），2021年，第6024–6028页。[36] X. Shi、P. Zhou、W. Chen 和 L. Xie，《Darts-Conformer：迈向高效的端到端自动语音识别基于梯度的神经架构搜索》，*arXiv* 预印本 *arXiv:2104.02868*，2021年。[37] S. Kim 等，《Q-ASR：用于高效语音识别的纯整数零样本量化》，*arXiv* 预印本 *arXiv:2103.16827*，2021年。[38] M. Ravanelli 等，《SpeechBrain：通用语音工具包》。2021年。[39] E. Bakhturina、V. Lavrukhin 和 B. Ginsburg，《用于语音数据集构建的 NeMo 工具箱》，*arXiv* 预印本 *arXiv:2104.04896*，2021年。[40] D. Bermuth、A. Poeppel 和 W. Reif，《Scribosermo：适用于德语和其他语言的快速语音转文本模型》，*arXiv* 预印本 *arXiv:2110 07982*，2021年。.[41] M. A. Ulasik、M. Hürlimann、F. Germann、E. Gedik、F. Benites 和 M. Cieliebak，《CEASR：用于评估自动语音识别的语料库》，载于第十二届语言资源与评测会议，2020年，第6477–6485页。[42] Y. Sinha 和 I. Siegert，《通过结合不同的在线自动语音识别 API 输出，提高德语语音助手对话的识别准确率》，《口语人机交互的人类视角》，第11–16页，2021年。

[43] A. Lopez、A. Liesenfeld 和 M. Dingemanse，《荷兰语、英语和德语会话语音自动语音识别评估：缺失了什么？》，发表于 第*18*届自然语言处理会议（*KONVENS 2022*）论文集，2022 年，第 135–143 页。