arXiv:2301.06475v1 [cs.CL] 16 Jan 2023

使用 Kaldi 对奥地利德语会话语音进行自动语音识别

Julian Linke、Saskia Wepner、Gernot Kubin、Barbara Schuppler

信号处理与语音通信实验室，格拉茨工业大学，
Inffeldgasse 16c，8010 格拉茨，奥地利

摘要

随着对话系统变得越来越具互动性和社交性，对会话语音进行准确的自动语音识别（ASR）也日益重要。这使关注重点从简短、自发的任务导向对话转向复杂度高得多的日常面对面对话。然而，收集和标注此类对话耗时很长，而且针对这种特定说话风格的数据十分稀缺。本文介绍了以奥地利德语朗读语音和会话语音为目标的自动语音识别实验。为应对会话德语可用资源有限的问题，同时考虑说话人之间在发音特征方面存在较大差异，我们在基于 Kaldi 的自动语音识别系统中加入了一个（大型）基于知识的发音词典，并探索了不同的数据驱动方法，以限制每个词条的发音变体数量。我们在奥地利德语朗读语音上取得了 0.4% 的最佳词错误率，在会话语音上取得了 48.5% 的最佳平均词错误率。研究发现，使用我们的最佳发音词典所达到的性能，与将语言模型使用的数据量增加约 360% 至 760% 所达到的性能相近。研究结果表明，在低资源场景下，尽管语音技术总体上趋向于仅使用数据驱动方法，基于知识的方法仍是一种有效且高效的方案。

索引词——自动语音识别、会话语音、发音词典、低资源、奥地利德语

1. 引言

近年来，会话语音（CS）的自动语音识别（ASR）变得愈发重要。一方面，从语音技术的角度来看，人们对能够实现人机交互的社交机器人日益感兴趣。旨在充当对话伙伴的语音代理，只有在能够流畅、自然地进行对话时才有用，这包括掌控对话时机、处理说话者不流利现象，或进行恰当的假设重排序 [1]。另一方面，从语音科学的角度来看，人们也越来越关注对日常语音过程的语音学和心理语言学分析 [2, 3]。与受控发音实验中的语音相比，自发语音的变化更为丰富，因此，分析这种说话风格也需要更多语音数据，理想情况下，这些数据应通过自动语音识别系统自动转写（例如， [4]）。然而，会话语音的自动转写结果与人工标注员的转写文本仍有很大差异 [5]。尽管语音科学和语音技术领域都对会话语音抱有兴趣，但大多数最先进的自动语音识别系统仍然是为既非完全自发、也非互动式的说话风格而开发的。造成这一现象的原因之一，是缺乏足够的大规模（已转写或原始）语音资源来支持更自发的说话风格。本文直面低资源语言变体（奥地利德语）中随意、会话式语音的自动语音识别这一挑战。

1.1 自发语音语料库的分类

在自动语音识别文献中，说话风格的定义依据各不相同，而用于语料库分类的“朗读”“自发”和“会话式”等术语，实际上可能指向特征迥异的语料库。图1展示了一种分类方案，有助于我们描述本文所用语料库的风格；一般而言，也有助于我们确定自动语音识别社区广泛使用的哪些语料库实际上彼此具有可比性。请注意，图1并未呈现所有可能情况的完整分类，而是在本文所用语料库所属的节点处结束分类树。

一般而言，我们可以区分朗读语音和自发语音。自发语音与朗读语音的不同之处在于，词位及其词序是即兴规划的。朗读语音的例子包括 LibriSpeech [6], ，在该语料库上，最先进的语音识别系统的字错误率（WER）可达 1.4% [7]。此外，本文使用的基尔语料库 [8] 和 GRASS 语料库 [9]中的朗读语音部分也属于这一类别（更多详情见第2节）。

接下来，我们根据自发语音是否具有任务导向性加以区分。任务导向对话中的语音通常围绕特定领域展开，而且说话人不像在对话主题开放时那样随意交谈。任务导向对话还可以根据是否有实验人员在场进一步区分。我们将录音中负责引导对话的在场人员定义为实验人员。此人可能是语言学家，也可能是受过专业训练的广播采访者。这类语音的特点是话语的句法结构相对完整，且发音谨慎，这是因为参与者是受过训练的说话人。相比之下，在非正式对话中，说话人会大幅弱化发音，并产出句法不完整的结构[10]。没有实验人员在场的任务导向对话，例如 Verbmobil [11] 和基尔语料库 [8]中的对话。这些简短对话每段约持续 2 至 20 分钟，这样的时长不足以让说话人忘却录音情境，从而影响了最终说话风格的自然度。

另一种对语音语料库进行分类的方式，是依据交谈者的数量。我们将（自发）独白和面向机器的对话与两名或更多说话人之间的对话区分开来；前两者都不包含串话，后者例如包括轻松场合下三名说话人之间的对话 [12]，以及会议场景中更多说话人之间的对话 [13]。说话人数量越多，自动语音识别面临的挑战就越大，因为系统需要处理重叠语音，而这不仅带来声学上的困难，还涉及共同补全、话轮竞争和中断话轮等结构性语音现象。AMI会议语料库的词错误率约为 21.2% [14]。

聚焦于两名说话人之间的对话时，我们还可以进一步区分对话主题是轻松随意还是专业正式，因为我们认为，轻松随意的话题也会带来一种说话风格，其特点是发音变异更多，和／或方言变体使用得更频繁。当说话人彼此关系亲近（甚至可能来自同一方言区）时，发音变异可能会更加显著。在某些语言区域（例如奥地利德语）中，这些风格影响可能是连续的；而在另一些区域（例如瑞士德语）中，则可能呈现双言制 [15]。我们注意到，从社会语言学角度来看，关系对说话风格的影响并不能跨语言直接比较。

最后，我们根据会话语音是否发生在面对面的场景中进行分类。在广泛使用的 Switchboard语料库中，[16], 彼此熟识的说话人通过电话进行对话，无法利用视觉线索，还需要应对较低的音质，这迫使他们更加清晰地发音并避免说话重叠。Switchboard的自动语音识别结果的词错误率在4.3%至5.1%之间 [17, 18]。

来自斯图加特的 IMS GECO 数据库包含两名关系疏远的说话人之间的对话：面对面场景（GECO Multi），以及非面对面场景；在后者中，说话人被一堵实体墙隔开（GECO Mono）[19, 20]。使用 GECO 得到的首次词语识别结果仅正确识别了25%的词语 [21]。

本文关注的语料库是 GRASS 语料库中的会话式部分 [22],，其中包含两位关系亲近者之间的话题开放、轻松随意的面对面对话，每次持续整整一小时，且没有实验人员在场。目前，这种特定说话风格的可用数据仍然很少。日常日语会话语料库（CEJC） [23] 目前正在建设中，其中还将收录采用个体录音方法录制的材料。该材料包括 40 位受访者的录音，受访者在性别和年龄方面均衡分布；每段录音都将由便携式录音设备采集，在约两到三个月期间、各种日常情境下共收集 15 小时语音数据。据我们所知，目前尚未发表使用 CEJC 进行的自动语音识别实验；一项使用 CEJC 开展对话情境识别的研究显示，系统的结果尚未达到人工评估的水平 [24]。总而言之，CEJC 和本研究所用数据（GRASS CS）都包含由会话语音中的说话者互动所带来的多种挑战。

1.2. 本文目标

本文的主要目标是介绍针对 GRASS CS 的自动语音识别实验。GRASS CS 是一个奥地利德语会话语音语料库，共包含 30 小时的语音。目前还没有其他风格相近、来自同一语言变体的德语数据库。我们采用基于深度神经网络—隐马尔可夫模型的 Kaldi 模型 [27, 28] ，并测试使用其他德语语料库进行声学模型和语言建模是否能提升性能。鉴于德语与奥地利德语在发音上差异显著，我们探究将语言学知识融入发音词典在多大程度上能够提升自动语音识别性能。

面对奥地利德语会话语音，人们或许会问：自动语音识别的难点究竟在于语言变体（即德语与奥地利德语之间的差异），还是在于说话风格。为了区分这两种因素，我们也报告了朗读奥地利德语的自动语音识别结果，并表明，即使采用相当简单的 Kaldi 配置方案，也已能取得最先进的结果。

2. 材料

2.1. GRASS 语料库

格拉茨朗读与自发语音语料库（GRASS 语料库） [9, 22] 包含约 30 小时的奥地利德语朗读语音（RS）和会话语音（CS），由 38 位奥地利说话人（19 位女性、19 位男性）录制。由于会话语音中的语言使用会因教育程度、社会背景

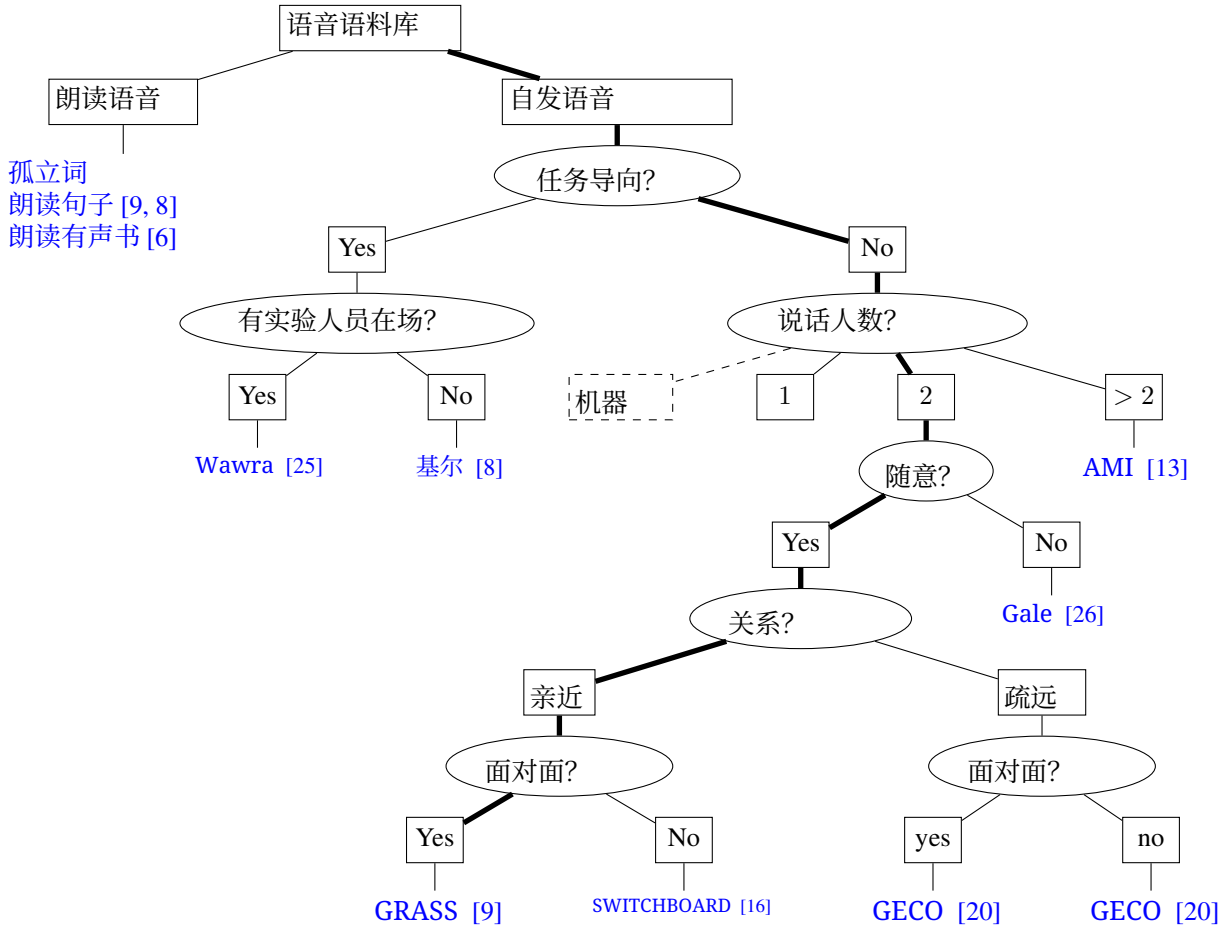

图1：不同说话风格的语音语料库分类。树形结构由陈述句（黑色矩形）、疑问句（椭圆）和选定示例（蓝色）构成。虚线矩形表示说话人与机器之间进行面向机器交互的可能性（即对话系统）。粗线表示 GRASS CS 的路径。

和方言区而有很大差异，因此说话人均选自同一大方言区（奥地利东部），出生于该地区，在城市居住多年，并拥有高等教育学历。录制 CS 部分时，研究人员选取了 19 对彼此相识多年的说话人，每对录制一小时，中间不间断，以促成流畅、自发的交谈。录音室内没有实验人员，也不限制话题选择或说话方式，因此录音中出现了自然、带有一定方言色彩的发音，以及重叠语音频繁、笑声和使用脏话等典型特征 [22]。尽管说话人知道自己正在被录音，但五到十分钟后，他们似乎完全忘记了身处录音室，开始随意交谈。完成一小时的 CS 录制后，说话人才朗读短篇故事和选定的独立句子。RS 和 CS 部分均由同一批说话人录制。

2.2. 基尔语料库

基尔德语口语语料库 [8] 共包含 5 小时的朗读语音与自发语音，主要由来自德国北部的说话人录制。该语料库提供两个自发语音部分：（1）“预约场景”部分，包含 43 位说话人（22 位女性／21 位男性）进行预约的约 4 小时对话。在这一场景中，只有参与者按住按钮时才会录制语音，而按住按钮的同时也会阻断对话对象的通道。因此，该场景有效避免了重叠语音。（2）“视频任务场景”部分包含约 1 小时的双人对话。在这一场景中，分别向两位参与者播放经过剪辑的电视剧片段，并要求他们找出视频中的差异。我们在奥地利会话语音实验中使用了基尔语料库（KICS）中的自发语音部分。

表1：所用数据集概览：德语和奥地利德语语料库，包含朗读语音（RS）和会话语音（CS）。

缩写	语种变体	小时数
GRASS CS	奥地利德语	14.3
GRASS RS	奥地利德语	4.6
GECO	德语	18.7
KIEL CS	德语	5

2.3. IMSGECO 数据库

IMS GECO 数据库（GECO）[19, 20] 包含 46 段自发对话，每段约 25 分钟，由彼此不熟悉的女性说话人在两种设置下进行：1）包含 22 段对话的单模态设置（GECO-Mono），参与者被实体墙隔开，因此无法看到彼此；2）包含 24 段对话的多模态设置（GECO-Multi），采用面对面对话形式，与 GRASS 会话语音相近。单模态设置涉及 12 位说话人，其中 7 位也参加了多模态设置，因此部分对话组合同时出现在 GECO-Mono 和 GECO-Multi 中。在这两种设置下，说话人都可以自由选择讨论主题。

3. 朗读语音的自动语音识别

3.1. 方法

声学模型（AM）和语言模型（LM）使用 RS 语料库中的数据进行训练。GRASS RS 数据集包含 6 小时的语音，每位说话人朗读的基本上都是相同的语音平衡句。训练集包含 33 位说话人（5.25 小时），验证集包含 2 位说话人（0.37 小时），测试集也包含 2 位说话人（0.34 小时）。由于一位说话人的朗读流利度异常不佳，我们将其排除。

我们提取了 13 维梅尔频率倒谱系数，并进行了倒谱均值和方差归一化（CMVN），同时比较了不同帧长 {20 ms、25ms、30ms} 以及帧移 {7.5 ms、10 ms、12.5 ms}的组合。对于声学模型（AM），初始的对角高斯混合模型-隐马尔可夫模型（短 GMM-HMM）包含使用梅尔频率倒谱系数+Δ+ΔΔ 特征进行的基础单音素和三音素训练。

词典是使用适用于标准德语的字素转音素在线工具 [29] 构建的。由于该资源不适用于德语的奥地利变体，我们在音素层面应用了一组规则，将其输出调整为符合标准奥地利德语发音 [30], 和音系弱化现象的形式，例如弱央元音删除。为了提高识别性能，我们通过三条规则缩减了字素转音素生成的音素集：（R1）替换规则，将所有齿龈擦音、后齿龈擦音和塞擦音清化（这是标准奥地利德语中常见的音系过程）；（R2）将双元音拆分为两个独立音素的规则；

表 2：GRASS RS 的最佳词错误率汇总。我们训练了一个高斯混合模型（见第 3.1 节），并比较了不同音素集简化规则对验证集和测试集的影响。

音素集规则	验证集	*test*
–	0.96	1.2
+R1	1.05	1.04
+R1+R2	**0.4**	**0.64**
+R1+R2+R3	0.56	**0.64**

（R3）根据奥地利德语的语音学研究 [31]，将短元音和长元音合并的规则。总之，我们将音素集从最初的 64 个音素缩减至 38 个。

我们使用 SRILM 工具包对不同阶数的 N 元语言模型（语言模型）采用 Witten-Bell 折扣[32]。语言模型根据整个朗读语音组件（训练集、验证集和测试集）中所有话语的文本生成，因为朗读语音组件中的所有说话人都朗读相同的文本。

3.2. 结果

表 2 总结了朗读语音的结果。首先，我们使用三元语言模型分析了不同帧移和帧长的影响。实验表明，与帧移 f_{sh} ={7.5 ms、10 ms 12.5 ms} 相比，帧长分别为 f_{len} = {20ms、25ms、30ms} 对词错误率的影响较小。使用 $f_{\text{sh}} = 12.5$ms 和 $f_{\text{len}} = 20$ ms 时，三音素词错误率最低（0.96 %）。单音素模型和三音素模型在大多数情况下表现相近。然而，当帧移为 $f_{\text{sh}} = 10$ms，帧长为 f_{len} = {25 ms、30 ms} 时，三音素模型的结果更差。为了进一步优化声学模型集合，我们依次应用了规则 R1、R2 和 R3。将最终词错误率与朗读语音部分进行比较（见表 2），R1 和 R2 使最佳三音素词错误率降低了 0.56 %。R3 则使结果略微变差，增加了 0.16 %。

接下来，我们使用最佳帧移 $f_{\text{sh}} = 12.5$ 毫秒和帧长 f_{len} =20 毫秒的配置，测试了不同阶数的语言模型 $\{1, 2, 3, 4, 5\}$。对于单音素模型和三音素模型而言，二元语法、三元语法、四元语法和五元语法的表现都相近（词错误率 $\approx 1 \ldots 2$ %）。我们决定采用三元语法模型，因为它略有优势，取得了最佳词错误率 $= 0.96$%。使用一元语法时，我们得到的结果无法进行比较，因为最佳词错误率差异很大。在这种情况下，单音素模型的最佳词错误率（词错误率 $= 35.77$ %）也明显差于三音素模型的最佳词错误率（词错误率 $= 18.25$%）。

3.3. 结论：朗读语音

我们针对朗读语音开展的自动语音识别实验表明，使用带有规范发音的词典（即不含发音变体）时，取得了最低的词错误率。对该规范词典唯一进行的调整，是根据奥地利标准德语的发音对音素集进行精简（例如，取消齿龈擦音的浊化）。使用基础三音素模型时达到了最先进的性能（测试集词错误率为 0.64）。我们观察到，所采用的方法选择带来了显著改进（例如帧移、音素集精简、声学模型遍历次数和语言模型阶数）。此外，有效集与测试集之间的词错误率差异相对较小（$\approx 0.01 - 0.24$ %）。总体而言，词错误率处于朗读语音其他先进系统的范围内（例如， [33] 中为 1.4% ）。

4. 会话语音自动语音识别

4.1. 方法

图2展示了实验设置的示意概览。声学模型和语言模型使用 GRASS CS 、 GECO 和 KIEL CS 的数据进行训练（见第 2 节）。我们介绍了分别仅使用 GRASS CS、使用 GRASS CS 和 GECO、使用 GRASS CS 和 KIEL CS，以及使用 GRASS CS、 GECO 和 KIEL CS 进行训练的实验。本研究通过针对 GRASS CS 进行留一p交叉验证（每组包含来自同一段对话的 $p = 2$ 位说话人），重点评估奥地利德语会话语音的自动语音识别。每个划分约有 0.8 小时的测试数据和 13.5 小时的训练数据。我们从所得训练划分中随机选取 10% 作为验证集（约 1.35 小时），用于调整基本模型参数。加入来自 GECO 或 KIEL CS 的训练数据时，则从新加入的训练数据中随机选取 10% 构建验证集。评估时，我们比较了由上述交叉验证得到的各个测试划分上的性能。

在 GRASS CS 预处理过程中，我们排除了包含笑声、歌唱、模仿声/拟声词、完全无法理解的词语标记或伪迹（例如说话者意外碰到麦克风）的片段。对于 GECO，我们从转写文本中删除了表示笑声、清嗓声和不完整词的符号。对于 KICS，我们从转写文本中删除了表示笑声、咂嘴声、不同类型噪声和重复的符号。

对于声学模型，自动语音识别的单音素和三音素训练步骤在大部分方面与朗读语音实验相似。首先，再次使用 13 维梅尔频率倒谱系数和倒谱均值和方差归一化训练两个模型（窗口分别为 $f_{\mathrm{sh}} = 10$ 毫秒和 f_{len} =25 毫秒）。在三音素高斯混合模型之上（见第 3.1 节），训练了一个采用线性判别分析和最大似然线性变换的说话人无关高斯混合模型，得到 GMM+LDA+MLLT。说话人自适应训练也在该模型之上进行 GMM+LDA+MLLT，并采用受约束最大似然线性回归（特征空间最大似然线性回归）进行训练，得到 GMM+fMLLR。使用最终的三音素对齐训练基线深度神经网络-隐马尔可夫模型混合模型，该模型由一个包含 13 层、隐藏维度为 512 的时延神经网络构成，仅使用预先计算好的梅尔频率倒谱系数特征。网络采用基于交叉熵准则的帧级目标函数进行训练。我们的方案基于 [36] 中发表的方案，相关的深度神经网络设置见 [37, 38]。

对于语言模型，我们使用 SRILM 工具包 [32] ，采用与朗读语音相同的配置，但训练的是三元语法。此处，语言模型根据 GRASS CS 交叉验证训练划分中所有话语的文本生成；如果另外两个德语语料库也用于声学模型训练，则也使用它们的文本。为了评估训练语言模型时可能存在的数据不足问题，我们还进行了实验，通过加入约 22 万条奥地利公共电视服务（AGS）面向聋人和听障人士的广播字幕中的奥地利德语句子，构建规模更大的三元语法模型 [39]。对于后者，我们通过再次加入从 AGS、德语维基百科[1] 和欧洲议会[2]中随机选取的 500 万条德语句子，使用四元语法进行语言模型重打分。为获得额外的语言模型数据，我们改造了 [40]中介绍的一款工具包。

4.2. 词典生成

我们根据所有语料库的转写文本创建了词表。规范发音由字素到音素在线工具 [41]生成。我们创建了四种不同的发音词典。

标准词典。由于所用工具的德语语言设置会生成德国标准德语 (GSG) 的发音[3], ，我们应用了 6 条输入切换规则，以获得奥地利标准德语的发音。我们将由此生成的发音词典称为标准词典。对于外语词，我们将语言设置更改为相应的语言。

allPVs。我们根据 [30] 的研究发现，对规范的奥地利德语发音应用 26 条音系规则，创建了包含发音变体（PVs）的词典。这些规则中有 17 条适用于所有德语变体的会话语音，例如塞音同化和音节韵尾中的 r 音删除；另外 9 条规则涵盖奥地利德语变体的典型发音，例如特定过去分词开头音节“ge”的删除。除了这些基于规则的变体，我们还手动创建了一些 PV，以涵盖无法通过自动方式生成、但在奥地利德语自发语音中很常见的发音，例如单词“wir”的发音 [maːˑ] ，其规范发音为

[1] https://dumps.wikimedia.org/dewiki/20220701 [2] https://www.statmt.org/europarl/v7/de-en.tgz [3] 德国标准德语指德国人所说的德语

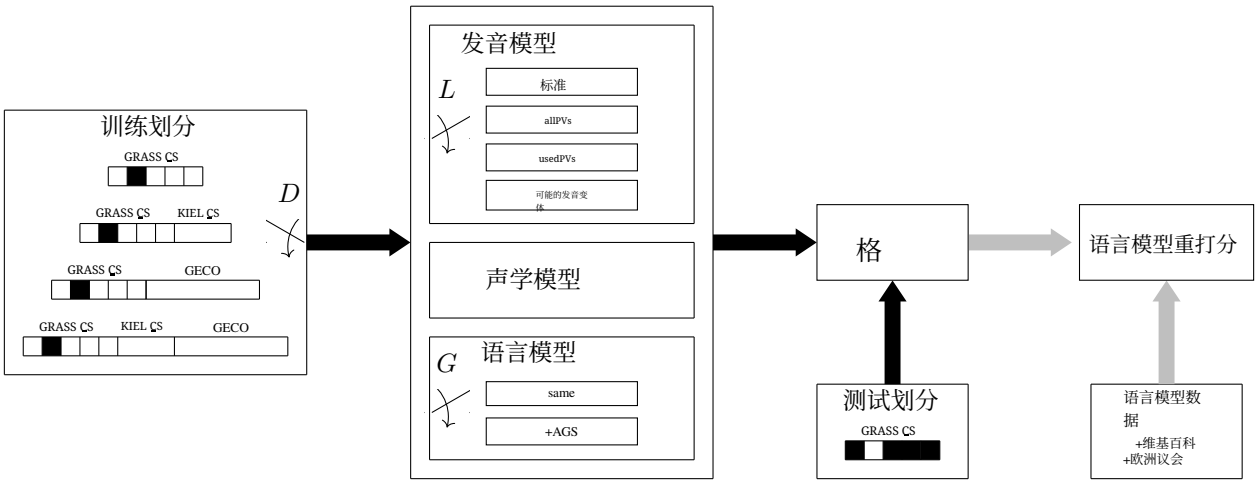

图2：会话语音识别实验的示意架构。训练基于 4×19数据组合，因为我们引入了4种可能的语料库组合（切换 D），并且对 GRASS CS 中的每段对话分别进行处理。由此得到4种数据组合下的19个依赖对话的训练和测试划分（见表4）。分析词典影响的实验（切换 L）对声学模型和语言模型训练使用相同的数据。分析语言模型影响的实验（切换 G）使用我们的标准词典，并将语言模型重打分作为额外选项（灰色箭头）。

[viːɒ]。最终得到的词典（allPVs）平均每个词包含 $5.57-6.18$ 个变体。

usedPVs。我们使用 allPVs 词典，通过强制对齐对所有语料库（即 GRASS CS、KIEL CS 和 GECO）进行分段（帧移为 f_{sh} =7.5 毫秒）。根据这些分段，我们提取说话人实际发出的发音变体，并仅使用这些发音变体创建发音词典。所得词典（usedPVs）中每个词平均有$1.37-1.43$ 个变体。

可能的发音变体。受 [42] 中所述方法的启发，我们创建了一个词典，其中只包含在强制对齐中出现频率较高的变体。与 [42]，一样，我们计算了发音概率估计所需的统计数据，但没有为特定发音纳入不同概率，而是只考虑估计概率为 $p > 0.65$ 的发音。这样选择是为了更好地与其他词典进行比较，因为在词典中引入额外的发音概率会改变实验设计（换言之，根据定义，所有最终词典中的发音变体概率都相同）。生成的词典（可能的发音变体）平均每个词包含 $1.16-1.26$ 个变体。表3汇总了自动语音识别实验中使用的所有不同发音词典。

4.3. 结果

表4展示了不同训练设置下的自动语音识别结果，测试数据始终采用 GRASS CS。我们比较了自动语音识别

表3：用于会话语音自动语音识别实验的不同发音词典条目统计（见第4.2节）。条目数受用于声学模型/语言模型训练的语料库数量影响（见表4或图2）。

词典名称	最小（#条目）	最大（#条目）
标准	13.9k	22.7k
allPVs	74.2k	135.1k
usedPVs	17.4k	30.4k
可能的发音变体	14.6k	26.9k

实验内容包括：(1) 使用不同规模的数据进行声学模型训练；(2) 使用包含不同数量变体的词典；(3) 使用不同规模的数据训练语言模型；(4) 组合最佳声学模型、词典和语言模型。就声学模型（AM）而言，所有声学模型都受益于加入其他语料库的训练数据；按均值计算，词错误率绝对值约降低了 $1-2\%$。不过，使用更多数据时，相应的标准差也更高，这表明整体性能有所提升，但也出现了稳健性问题。就发音词典而言，与使用我们的标准词典相比，结果显示，包含大量变体的词典（即 allPVs 和 usedPVs）会导致性能下降。而使用 likelyPVs 词典时，性能则有所提升；该词典包含少量可能的发音变体，与使用标准词典的基线最佳均值相比，词错误率约降低了 1.5%。至于调整

表4：使用我们的 Kaldi 基线系统获得的奥地利德语会话语音依赖对话的词错误率汇总。前两列显示用于声学模型和语言模型训练的数据，第三列显示所用词典，其余各列给出所得 19 个词错误率的均值和标准差，以及相应的最小和最大词错误率。

	AM	LM	词典	WERs	最小（词错误率）	最大（词错误率）
Baseline	GRASS CS	same	标准	53.89/5.18	42.1	63.7
	GRASS CS +KIEL CS			53.22/5.23	42.0	63.2
	GRASS CS +GECO			52.58/5.59	40.5	63.9
	GRASS 会话语音 +GECO+KIEL 会话语音			**51.91/5.74**	**39.7**	**63.2**
Influence Lexicon	GRASS 会话语音	same	allPVs	55.56/5.03	43.9	64.7
	GRASS 会话语音 +KIEL 会话语音			55.16/5.21	44.4	64.8
	GRASS 会话语音 +GECO			54.37/5.45	42.6	64.7
	GRASS 会话语音 +GECO+KIEL 会话语音			53.62/5.56	41.6	64.7
	GRASS CS	same	usedPVs	55.22/4.88	43.8	64.3
	GRASS CS +KIEL CS			55.06/5.15	43.5	64.5
	GRASS CS +GECO			54.15/5.4	42.5	64.7
	GRASS CS +GECO+基尔 会话语音			53.69/5.52	42.9	64.3
	GRASS CS	same	可能的发音变体	51.87/4.88	40.3	**60.9**
	GRASS CS +KIEL CS			51.64/5.22	40.0	61.9
	GRASS CS +GECO			50.93/5.61	38.8	62.2
	GRASS 会话语音 +GECO+基尔 会话语音			**50.48/5.66**	**39.7**	62.0
Influence LM	GRASS CS	+AGS (220k)	标准	52.26/5.5	40.1	62.9
	GRASS CS +KIEL CS			51.53/5.58	40.3	62.8
	GRASS CS +GECO			51.38/5.72	40.3	63.3
	GRASS 会话语音 +GECO+基尔 会话语音			50.74/5.82	39.0	62.3
	GRASS CS	+AGS (220k) +重打分 (5M)	标准	51.17/5.26	39.5	62.4
	GRASS CS +KIEL CS			50.91/5.65	39.8	62.4
	GRASS CS +GECO			50.92/5.82	39.5	63.2
	GRASS 会话语音 +GECO+基尔 会话语音			**50.19/5.82**	**38.7**	**62.2**
Best	GRASS CS	+AGS（220k） +重打分（5M）	可能的发音变体	49.15/5.28	37.3	**59.1**
	GRASS CS +KIEL CS			49.02/5.69	37.3	60.4
	GRASS CS +GECO			49.07/5.88	37.2	60.8
	GRASS 会话语音 +GECO+基尔 会话语音			**48.5/6.09**	**37.0**	61.3

语言模型的训练数据量，我们通过在声学模型训练中加入所有语料库的数据、在语言模型训练中加入所有语料库及 AGS 的数据、使用包含可能的发音变体的词典，并利用额外的 5M 个德语句子进行语言模型重打分（见第 2 节），取得了最佳结果，词错误率的最佳均值为 48.5%。

总体而言，与我们的基线系统相比，最佳词错误率均值绝对降低了约 4.5%。总的来说，在所有实验中，我们观察到 GRASS CS 不同对话（即说话人对）之间的词错误率差异很大（标准差介于 4.88% 至 6.09% 之间）。

4.4. 讨论

本文旨在构建一个基于 Kaldi 的奥地利德语自动语音识别系统，重点关注会话语音。由于初步实验已显示不同说话人组合之间存在显著差异，我们决定提供交叉验证结果，以便更深入地了解自动语音识别系统的会话依赖性。值得注意的是，尽管通过某些方法选择取得了性能提升，我们仍然观察到不同对话之间的词错误率存在同样高的标准差。因此，无论是改变声学模型和语言模型训练所用的数据规模，还是采用不同的发音建模方法，都未能使自动语音识别系统对会话语音的变异更具鲁棒性 [43]。

与其他基准测试相比，我们采用交叉验证方法所得的结果凸显了对话语音识别任务的挑战性。其他基准测试往往使用预先定义的训练集和测试集进行训练和测试，这可能会导致自动语音识别准确率出现乐观偏差 [44]。例如， [45], 中所述的跨方言分析显示，当用某一方言训练的系统评估另一种同一语言方言时，自动语音识别性能会下降。我们推测，逐一测试每段对话会呈现出类似的效应，因为尽管 GRASS 中的说话人地域背景相近，我们仍发现了显著的个体方言差异，这与此前对该语料库的分析结果一致 [30]。我们的结果表明，不仅应研究如何提升整体自动语音识别性能，也应更加着力于弥补鲁棒性不足，尤其是在会话语音识别中。

使用我们的标准词典和大量额外语言模型数据进行自动语音识别，词错误率均值为50.19%。仅使用包含可能发音变体的词典（在标准词典中增加约4.2千个词条；见表3），而不增加大量额外语言模型数据时，我们的词错误率均值为50.48%。因此，将使用不同词典得到的结果与使用不同数量语言模型数据得到的结果进行比较（参见表4），我们发现，使用大量数据训练语言模型对降低词错误率的效果同样显著（约1.5%），与使用最佳发音词典（即可能的发音变体）不相上下。一项关于自动语音识别发音变异建模的综述 [46]总结指出，将发音变体加入词典似乎能够提升识别性能，尤其是在考虑到不同变体的出现频率时。二十年过去，自动语音识别架构也经历了诸多发展，我们的结果依然证实了这一观察。我们还表明，成熟的发音建模无需额外的数据资源，也不需要很大的计算开销，便可弥补收集更多语言模型数据的必要性。然而，将这两种方法（发音建模和收集更多语言模型数据）结合起来，仍能取得最佳结果。这一发现对低资源自动语音识别领域尤为重要。

5. 结论

本文介绍了一个用于奥地利德语朗读语音（RS）和会话语音（CS）语料库的自动语音识别系统的开发，着重应对两项挑战：（1）录制的会话语音非常随意；（2）奥地利德语是德语的一种变体，通常可用的语音数据很少，而会话语音的数据更是如此。对于RS，使用 $f_{\mathrm{sh}} = 12.5$毫秒、规范词典、精简音素集以及高于 $N = 2$的语言模型阶数时，我们取得了最佳结果。对于CS基线系统，使用混合深度神经网络—隐马尔可夫模型，以及包含最可能发音变体的词典和一个基于德语和奥地利德语自发语音组合训练的语言模型时，我们取得了最佳结果。无论是将知识融入发音词典，还是扩充训练数据，都能带来相近的性能提升。我们观察到，不同对话之间的性能差异很大（即，标准差约为 $5–6$%），且这一现象与整体性能无关，表明该自动语音识别系统对会话语音的鲁棒性较低。鲁棒性不足可能源于多方面的显著差异，包括发音变异（即方言背景）、语速（会话语音的语速介于每秒0.88至45.45个音素之间，均值为12.38，标准差为4.28）、词汇选择（因为话题由说话人自由选择）、说话人是否使用完整的句法结构，以及其轮流发言行为。未来，我们计划详细分析阻碍会话语音鲁棒性自动语音识别的因素。

6. 致谢

Saskia Wepner 的研究工作由奥地利科学基金（FWF）资助，资助编号为 P-32700-N。

7. 参考文献

[1]T. Baumann、C. Kennington、J. Hough 和 D. Schlangen，《识别会话语音：增量式自动语音识别应为对话系统做什么，以及如何实现》，载于 IWSDS 会议论文集，2016年，第 1–12 页。[2] Benjamin V. Tucker 和 Mirjam Ernestus，《为什么我们需要研究随意语音，才能真正理解语言的产生、加工和心理词库》，《心理词库》，第 11 卷，第 3 期，第 375–400 页，2016年。[3] Petra Wagner、Jurgen Trouvain 和 Franz Zimmerer，《为语音研究中的风格多样性辩护》，《语音学杂志》，第 48 卷，第 1–12 页，2015年。[4] F. Cangemi、M. Clayards、O. Niebuhr、B. Schuppler 和 M. Zellers（编），《重新思考弱化：语音变异的条件、机制和领域的跨学科视角》，Mouton de Gruyter，柏林，2017年。[5] Alianda Lopez、Andreas Liesenfeld 和 Mark Dingemanse，《荷兰语、英语和德语会话语音自动语音识别的评估：遗漏了什么？》，载于第 18 届自然语言处理会议（KONVENS 2022）论文集，2022年，第 135–143 页。

[6] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，《LibriSpeech：基于公有领域有声书的自动语音识别语料库》，载于 ICASSP 会议论文集，2015年，第 5206–5210 页。

[7] Yu Zhang、James Qin、Daniel S. Park、Wei Han、Chung- Cheng Chiu、Ruoming Pang、Quoc V. Le 和 Yonghui Wu，《推进半监督学习在自动语音识别中的极限》，2020年。

[8] K. J. Kohler、B. Peters 和 M. Scheffers，《基尔德语口语语料库——朗读语音与自发语音：新版，修订并扩充》，2017年。

[9] B. Schuppler、M. Hagm¨uller、J. A. Morales-Cordovilla 和 H. Pessentheiner，《GRASS：格拉茨朗读与自发语音语料库》，发表于 LREC，2014年，第1465–1470页。

[10]K. Johnson，《美式英语会话语音中的大幅缩减》，发表于《自发语音：数据与分析》会议论文集。第十届国际研讨会第一场会议论文集，2004年，第29–54页。

[11] Wolfgang Wahlster，《Verbmobil：面对面对话的翻译》，载于《第四届机器翻译峰会论文集》，日本神户，1993年7月19日至22日，第127–136页。

[12] Francisco Torreira、MartineAdda-Decker 和 Mirjam Ernestus，《奈梅亨随意法语语料库》，Speech Communication，第52卷第3期，第201页，2010年1月。

[13] Jean Carletta、Simone Ashby、Sebastien Bourban、Mike Flynn、Mael Guillemot、ThomasHain、Jaroslav Kadlec、Vasilis Karaiskos、Wessel Kraaij、Melissa Kronenthal、Guillaume Lathoud、Mike Lincoln、Agnes Lisowska、Iain McCowan、Wilfried Post、Dennis Reidsma 和 Pierre Wellner，《AMI会议语料库：预发布》，载于《多模态交互机器学习》，Steve Renals 和 Samy Bengio 编，柏林、海德堡，2006年，第28–39页，Springer Berlin Heidelberg。

[14] Naoyuki Kanda、Guoli Ye、Yu Wu、Yashesh Gaur、Xi- aofei Wang、Zhong Meng、Zhuo Chen 和 Takuya Yoshioka，《利用单个疏远麦克风进行会议转录的端到端多说话人自动语音识别大规模预训练》，载于 Interspeech 2021 论文集，2021年，第3430–3434页。

[15] Agnieszka Ste.pkowska，《双言现象：瑞士案例批判性综述》，Studia Linguistica Universitatis Iagellonicae Cracoviensis，第129卷，第199–209页，2012年。

[16]J. J. Godfrey、E. C. Holliman 和 J. McDaniel，“SWITCHBOARD：用于研究与开发的电话语音语料库”，载于 ICASSP 会议论文集，1992 年，第 1 卷，第 517–520 页。

[17] Zolt´an T¨uske、George Saon 和 Brian Kingsbury，“英语会话语音识别的极限”，载于 Interspeech 2021 会议论文集，2021年，第 2062–2066 页。[18] W. Xiong、L. Wu、F. Alleva、J. Droppo、X. Huang 和 A. Stolcke，“微软 2017 年会话语音识别系统”，载于 ICASSP 会议论文集，2018年，第 5934–5938 页。[19] Antje Schweitzer 和 Natalie Lewandowski，“自发语音中发音速率的趋同”，载于 Interspeech 2013 会议论文集，2013年，第 525–529 页。

[20] Antje Schweitzer、Natalie Lewandowski、Daniel Duran 和 Grzegorz Dogil，《请注意！扩展 GECO 数据库》，载于 ICPhS，2015年。

[21] D. Arnold、F. Tomaschek、K. Sering、F. Lopez 和 R. H. Baayen，《自发会话语音中的词语可由一种错误驱动学习算法以近似人类的准确度识别；该算法可直接从智能声学特征中区分含义，绕过将音素作为识别单位》，PLOS ONE，第12卷，第4期，文章编号 e0174623，2017年。

[22] Barbara Schuppler、Martin Hagm¨uller 和 Alexander Zahrer，《一套朗读与会话式奥地利德语语料库》，Speech Communication，第 94 卷，62–74 页，2017年。

[23]Hanae Koiso、Yasuharu Den、Yuriko Iseki、Wakako Kashino、Yoshiko Kawabata、Ken’ya Nishikawa、Yayoi Tanaka 和 Yasuyuki Usuda，《日常日语会话语料库的构建：阶段性报告》，载于第十一届国际语言资源与评估会议论文集（LREC 2018），2018年。

[24] Yuya Chiba 和 Ryuichiro Higashinaka，《利用多模态信息识别日常会话的对话情境》，载于 Interspeech 2021 论文集，2021年，241–245 页。

[25] Daniela Wawra，《求职面试语料库：语料库语言学中的数据转录与主要议题》，Peter Lang Verlag，德国柏林，2014年。

[26] 等。Meghan、Glenn，《Gale 第二阶段阿拉伯语广播对话语音》，2013年，费城：语言数据联盟（LDC）。

[27]D. Povey、A. Ghoshal、G. Boulianne、L. Burget、O. Glembek、N. Goel、M. Hannemann、P. Motlicek、Y. Qian、P. Schwarz、J. Silovsky、G. Stemmer 和 K. Vesely，《Kaldi 语音识别工具包》，IEEE自动语音识别与理解研讨会论文集，2011年。

[28] Daniel Povey、Xiaohui Zhang 和 Sanjeev Khudan- pur，《采用自然梯度和参数平均的深度神经网络并行训练》，CoRR，第 abs/1410.7455 卷，2015年。[29] U. D. Reichel，《PermA 和 Balloon：字符串对齐与文本处理工具》，载于 Interspeech 论文集，2012年。[30]B. Schuppler、M. Adda-Decker 和 J. A. Morales- Cordovilla，《朗读及会话奥地利德语中的发音变异》，载于 Interspeech 论文集，2014年。[31] S. Moosm¨uller，《标准奥地利德语元音：声学语音学与音系分析》，大学任教资格论文，维也纳大学，2007年。[32] Andreas Stolcke，《Srilm——可扩展的语言建模工具包》，载于 Interspeech 2002 论文集，2002年。[33]Yu-An Chung、Yu Zhang、Wei Han、Chung-Cheng Chiu、James Qin、Ruoming Pang 和 Yonghui Wu，《w2v-bert：结合对比学习与掩码语言建模的自监督语音预训练》，2021年 IEEE自动语音识别与理解研讨会（ASRU），第244–250页，2021年。[34] R. A. Gopinath，《采用高斯分布进行分类的最大似然建模》，载于 ICASSP 论文集，1998年，第661–664页，第2卷。[35] M.J.F. Gales，《用于基于隐马尔可夫模型的语音识别的最大似然线性变换》，Computer Speech and Language，第12卷，第2期，第75–98页，1998年。[36]Josh Meyer，《Easy-kaldi》，https://github.com/JRMeyer/easy-kaldi，2020年。[37] Shakti P. Rath、Daniel Povey、Karel Vesel´y 和 Jan ˇCernock´y，《改进深度神经网络的特征处理》，载于 Interspeech 2013 论文集，2013年，第109–113页。[38]Karel Vesel´y、Martin Karafi´at 和 Frantiˇsek Gr´ezl，《用于大词汇量连续语音识别的卷积瓶颈网络特征》，载于 2011年 IEEE自动语音识别与理解研讨会，2011年，第42–47页。[39]《ORF-TVthek：面向聋人和听障人士的广播节目》，https://tvthek.orf.at/gehoerlosenserv
ice，
访问日期：2022-07-20。[40] Benjamin Milde 和 Arne K¨ohn，《德语开源自动语音识别》，载于 ITG 2018 论文集，德国奥尔登堡，2018年，第 251–255 页。[41] U. D. Reichel 和 T. Kisler，《作为网络服务的语言无关字素-音素转换与词重音分配》，《语言交流研究文本：Elektronische Sprachsignalverarbeitung 2014》，第 42–49 页，2014年。[42] Guoguo Chen、Hainan Xu、Minhua Wu、Daniel Povey 和 Sanjeev Khudanpur，《自动语音识别的发音与静音概率建模》，载于 Interspeech 2015 论文集，2015年，第 533–537 页。[43] Julian Linke、Philip N. Garner、Gernot Kubin 和 Barbara Schuppler，《对话语音识别需要数据吗？奥地利德语实验》，载于第十三届语言资源与评测会议论文集，法国马赛，2022年6月，第 4684–4691 页，欧洲语言资源协会。[44] Piotr Szyma´nski、Piotr ˙Zelasko、Mikolaj Morzy、Adrian Szymczak、Marzena ˙Zyła-Hoppe、Joanna Banaszczak、Lukasz Augustyniak、Jan Mizgajski 和 Yishay Carmiel，《我们实际的词错误率与我们以为的词错误率》，载于计算语言学协会论文集：EMNLP 2020 年会论文集。2020年11月，第 3290–3295 页，计算语言学协会。[45] Mohamed G. Elfeky、Pedro Moreno 和 Victor Soto，《多方言语言对语音识别的影响：选择太多可能有害》，《Procedia Computer Science》，第 128 卷，第 1–8 页，2018年，第一届自然语言与语音处理国际会议。[46] Helmer Strik 和 Catia Cucchiarini，《自动语音识别的发音变异建模：文献综述》，《Speech Communication》，第 29 卷，第 2 期，第 225–246 页，1999年。