arXiv:2305.01155v1 [eess.AS] 2 May 2023

MDPI

文章ATCO2项目经验总结：5000小时空中交通管制通信数据助力鲁棒自动语音识别与理解

Juan Zuluaga-Gomez *,[1,2] **Iuliia Nigmatulina**[1,3],**Amrutha Prasad**[1,4],**Petr Motlicek**[1,4],**Driss Khalil**[1],**Srikanth Madikeri**[1], **Allan Tart**[5], **Igor Szoke**[4], **Vincent Lenders**[6],**Mickael Rigault**[7],**Khalid Choukri**[7]

[1] Idiap研究所，瑞士马蒂尼[2] 洛桑联邦理工学院，瑞士[3] 苏黎世大学，瑞士[4] 布尔诺理工大学，捷克布尔诺[5] OpenSky Network，瑞士布格多夫[6] 网络防御园区，armasuisse，瑞士图恩3602* 通讯作者：juan-pablo.zuluaga@idiap.ch

摘要： 空中交通管制员与飞行员之间的语音通信对于确保空中交通管制安全高效至关重要。这项工作要求空中交通管制员保持高度警觉，而且可能枯燥乏味、容易出错。近年来，人们开始尝试将人工智能融入空中交通管制，以减轻空中交通管制员的工作负荷。然而，开发数据驱动的空中交通管制人工智能系统需要大规模的已标注数据集，而目前该领域恰恰缺少此类数据集。本文总结了ATCO2项目的经验。该项目旨在开发一个独特的平台，用于实时采集和预处理大量空域空中交通管制数据。音频和监视数据通过志愿者社区拥有的甚高频接收机，从公开可访问的无线电频率频道采集，随后上传至OpenSky Network服务器；这些服务器可被视为“无限数据源”。此外，本文还回顾了ATCO2合作伙伴以往的工作，包括：(i) 鲁棒自动语音识别，(ii) 自然语言处理，(iii) 空中交通管制通信的英语语言识别，以及(iv) 融合广播式自动相关监视等监视数据。我们相信，在*ATCO2*项目期间开发的处理流程，加上数据的开放共享，将推动空中交通管制领域的研究。ATCO2语料库的部分样本可在以下网站获取：https://www.atco2.org/data；完整语料库则可通过ELDA购买，网址为http://catalog.elra.info/en-us/repository/browse/ELRA-S0484/。我们证明，在几乎没有或完全没有空中交通管制领域内数据的情况下，ATCO2是开发自动语音识别引擎的合适数据集。例如，使用Kaldi的CNN-TDNNf模型，我们在公开空中交通管制数据集上取得了低至17.9%和24.9%的词错误率，分别比使用域外监督式CNN-TDNNf模型低6.6%和7.6%。我们希望本文能推动空中交通管制领域数据驱动人工智能系统的发展，并最终带来更安全、更高效的空中交通管制人工智能工具。

C 引用：ATCO2 的经验教训：5000 小时空中交通管制C通信，助力实现稳健的A自动语音识别与U 理解。航空航天 **2023**，*1*，0。https://doi.org/

收稿日期:
录用日期:
发布日期:

版权所有： © 2023 作者。本文已提交至航空航天，拟根据知识共享署名（CC BY）许可协议（https://creativecommons.org/licenses/by/ 4.0/）的条款和条件进行开放获取出版。

关键词： 空中交通管制通信；自动语音识别与理解；开放天空网络；呼号识别；广播式自动相关监视数据。

1. 引言

由于自动语音识别（ASR）及空中交通管制（ATC）语音理解系统有望提高航空业的安全性和效率，因此人们对其开发的兴趣日益浓厚。ASR 和理解系统的应用

空中交通管制中的技术推动了先进概念验证引擎的诞生，这些引擎可以协助空中交通管制员（空中交通管制员）完成日常任务。这些引擎旨在分析口语空中交通管制通信，并将其转换为机器可读文本，从而实现更快速、更准确的处理。MALORCA [1–4], HAAWAII [5,6], 或 StarFish [7] 等先前的研究表明，成熟度足够的引擎能够减轻空中交通管制员的工作负荷[1]，同时提高安全性。因此，自动语音识别和理解技术正变得日益先进，能够应对空中交通管制通信的复杂性，从而提升航空业的安全性和效率。主要的限制在于，所有这些引擎都是针对特定解决方案设计的，即某一个机场。将基于人工智能的引擎适配到不同机场或管制区域，需要新的领域内数据，这仍是一项挑战。例如，从一个机场（如机场 X）收集的空中交通管制音频数据通常无法成功迁移到另一个机场（如机场 Y）。

从事空中交通管制自动语音识别工作的社区所面临的最大瓶颈之一，是音频数据的收集与标注。这一过程成本高昂且要求严格：标注一小时的空中交通管制语音至少需要八小时的人力 [8,9]。对于某些解决方案，例如面向小型机场的解决方案，这一成本可能令人难以承受。因此，人们不禁要问，如何以最高效的方式收集和处理更多空中交通管制音频数据。幸运的是，ATCO2项目等近期项目旨在克服这一障碍 [10]。ATCO2语料库生态系统的总体概览见图1。

ATCO2项目旨在通过采用先进的自动语音识别和自然语言处理（NLP）系统，减少收集、预处理和转录空中交通管制语音通信所需的人力 [11,12]。ATCO2发布了迄今为止规模最大的空中交通管制语音通信语料库，其中包含超过5,000小时的自动转录音频数据及其对应的监视数据 [11]。此外，项目还发布了四小时的人工转录数据（即金标准标注）。通过ATCO2项目，我们发现，为数据标注员提供自动转录数据，而不是要求他们从头开始生成标注，可以大幅加快转录过程。这使标注过程的实时因子（RTF）降低了30个点，从50 RTF降至20 RTF [10]。

本文涵盖了与数据收集和标注处理流程、负责收集与标注的主要参与者，以及 ATCO2 期间开发的主要基于人工智能的系统相关的多个方面和经验教训。本文还介绍了空中交通管制相关自动语音识别与理解方面的基线系统和新技术，例如将监视数据整合到自动语音识别处理流程中 [5,13,14]。

本文其余部分的组织结构如下。第 2 节介绍了自动语音识别与理解相关的一些研究工作。第 3 节将介绍 ATCO2生态系统、数据采集流水线及其他一些特征。第 4 节将介绍数据收集平台（前端和后端）的技术细节，以及志愿者社区如何与其交互。第 5 节将介绍利用 ATCO2语料库可以开发的主要技术。最后，我们将在第 6 节简要介绍未来工作方向并总结全文。

2. 空中交通管制早期工作

近期关于空中交通管制通信的自动语音识别和理解的工作已记录于 [9,15]。近年来，更多面向研究的工作聚焦于纯自动语音识别。例如， [16] 建立了首个针对不同空中交通管制通信数据库的自动语音识别基准测试。此外，人们已投入大量精力，利用 [1,17,18] 等监视数据，整合新颖的半监督学习算法，以增强自动语音识别性能。

[1] 文献 [3]介绍了一个明确的示例。作者得出结论：集成新型基于自动语音识别的工具，可以使空中交通管制员在工作站中输入和确认许可指令所花费的总时间绝对减少 20 个百分点。

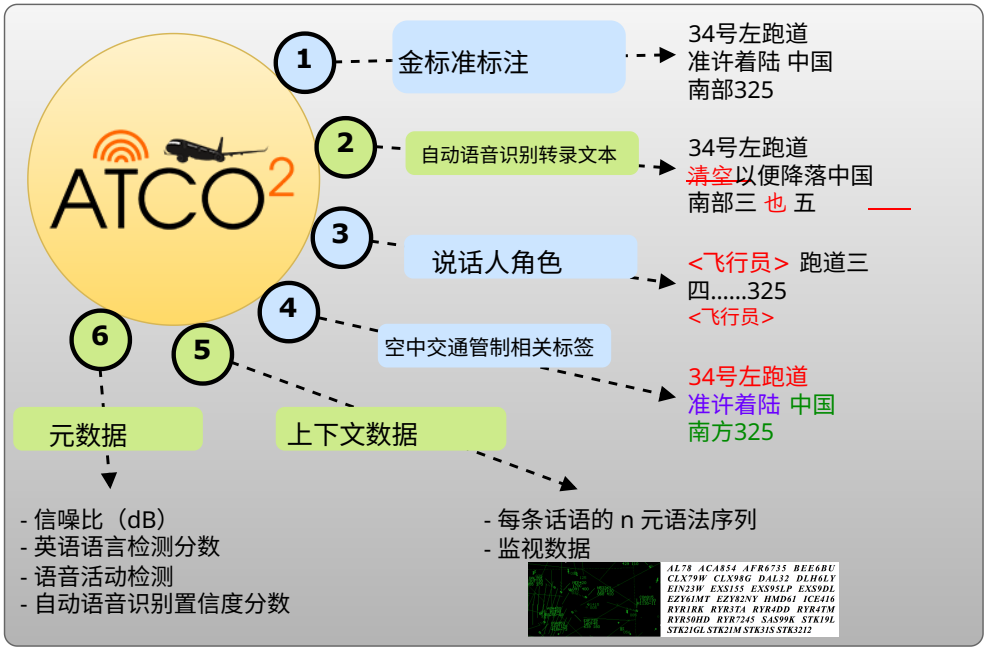


图1. *ATCO2*语料库生态系统。蓝色圆圈表示仅适用于*ATCO2*测试集语料库的注释。绿色圆圈表示适用于*ATCO2*测试集和*ATCO2*伪标注语料集的注释和元数据。摘自先前的研究 [10]。

这进一步印证了研究界对面向空中交通管制（ATC）的自动语音识别（ASR）及理解技术日益浓厚的兴趣，以及成熟的概念验证引擎能够协助空中交通管制员（ATCo）完成日常工作。我们此前开展的相关研究涉及从全球不同机场大规模自动采集空中交通管制音频数据，详见 [11,19]。此外，近期研究还探索了将监视数据整合到处理流程中，以提高呼号识别的准确性，详见 [5,13,14]。

另一项研究致力于开放 ATC 相关数据库：面向美国空中交通管制通信的数据库 [20]、捷克的数据库 [21],以及 [22]涵盖多种英语口音的数据库。最近，空中客车牵头举办了一项空中交通管制通信挑战赛，使用了来自法国的带法国口音的录音 [23,24]。研究人员也探索了私人数据库 [25,26]。有关空中交通管制相关数据库的概述，请参阅先前研究 [10] 中的表1；有关 ATCO2 项目发布的数据库，请参阅表1。

3. ATCO2语料库

众所周知，基于人工智能的工具在训练过程中需要大量可靠的转录数据。例如，如果我们拥有大规模数据，用于空中交通管制的自动语音识别或自然语言处理工具就能发挥更好的作用。ATCO2语料库旨在通过解决三大挑战，应对这一数据稀缺问题：

(1) 当前与空中交通管制相关的语料库主要侧重于自动语音识别。然而，要让基于人工智能的引擎成功部署到管制室中，就不能只准确转写空中交通管制通信，还必须理解其内容。这包括检测说话人角色（SRD），以及提取和解析呼号与指令的能力。ATCO2语料库通过提供用于说话人角色检测以及呼号和指令提取的详细标签，为这一挑战提供了全面的解决方案。这将进一步提升基于人工智能的系统在空中交通管制运行中的准确率和效率。

表1。 ATCO2项目发布的空中交通管制通信语料库。†去除静音后的完整数据库。††说话人的口音取决于机场所在地；由于隐私法规，在通信过程中的任何时刻均无法获知飞行员的口音。

数据库	详情	已授权	口音	小时†	Ref
	***ATCO2*项目发布的语料库**				
***ATCO2*语料库**	Data from 不同 机场 and 国家： 公共 语料库 catalog.elra.info/en-us/repository/browse/ELRA-S0484/		若干††		[10]
↪ *ATCO2* 测试集	用于自动语音识别和自然语言处理研究的真实数据。	✓	↪	4	[10]
↪ *ATCO2-PL* 数据集	用于研究的伪标注真实数据，涵盖自动语音识别和自然语	✓	↪	5281	[10,11]
	***ATCO2*项目发布的免费开放访问数据库**				
↪ *ATCO2-test-set-1h*	“自动语音识别数据集”：公开的 1 小时样本，a ATCO2 测试集的子集*ATCO2* 测试集。 https://www.atco2.org/data	✓	↪	1	[11]
↪ *ATCO2-ELD* 数据集	’英语语言检测数据集’：适用于英语的公开数据集 English 英语语言检测。 https://www.atco2.org/数据	✓	↪	26.5	[27]

(2) 域外的自动语音识别和基于自然语言处理的语料库难以迁移到空中交通管制领域。空中交通管制通信遵循独特的语法结构，并采用国际民用航空组织规定的一套专用词汇 [28], ，因此属于一个小众应用领域。这对域外语料库的使用构成了重大限制。[2] 因此，ATCO2项目收集并公开发布了大量空中交通管制专用数据，以助力开发面向空中交通管制的自动语音识别和理解引擎。

（3）从事空中交通管制研究的学术界受到公开可用的已标注数据严重匮乏的制约。为解决这一问题，ATCO2项目发布了一个包含超过5000小时自动标注数据（即*ATCO2-PLset*）的大型语料库，以及4小时的人工标注数据（即*ATCO2-test-set-4h*），如表1所示。[3]

*3.1. ATCO2*生态系统与概述

ATCO2生态系统如图所示 1。在收集 ATCO2语料库期间，我们遵循了若干预处理步骤，以规范生成的标注。我们的目标是尽量减少由语音不相似性导致的错误，例如 **“下降至两千英尺”** 和 **“下降，二二千英尺”**。我们按照国际民用航空组织的规则 [28] 以及广为人知的空中交通管制通信本体 [32]执行了若干文本规范化步骤，以统一人工标注和自动标注。标注协议概要如图所示 2。此外，关于数据标注中文本规范化和词典的更详细概述，请参阅 [10]第 3 节。ATCO2语料库由*ATCO2-PL* 集 语料库和 *ATCO2* 测试集 语料库组成，具体如下：

• 首先，***ATCO2-PL-set*语料库是首个专为 ∼5281小时、涵盖十座不同机场的空中交通管制语音的大规模数据集，见表2。据作者所知，

[2] 既往研究 [16] 表明，使用与空中交通管制无关的语料库（如 LibriSpeech [29]或其他语料库 [30,31]）无法匹配空中交通管制通信的声学特征，因此无助于自动语音识别训练。[3] 值得注意的是，事实证明，自动化工具生成的标注十分可靠；如 MALORCA 项目中的 Malorca-Vienna-test 集所示，仅使用这些转录文本训练自动语音识别引擎，词错误率（WER）可低至 9% [10]。

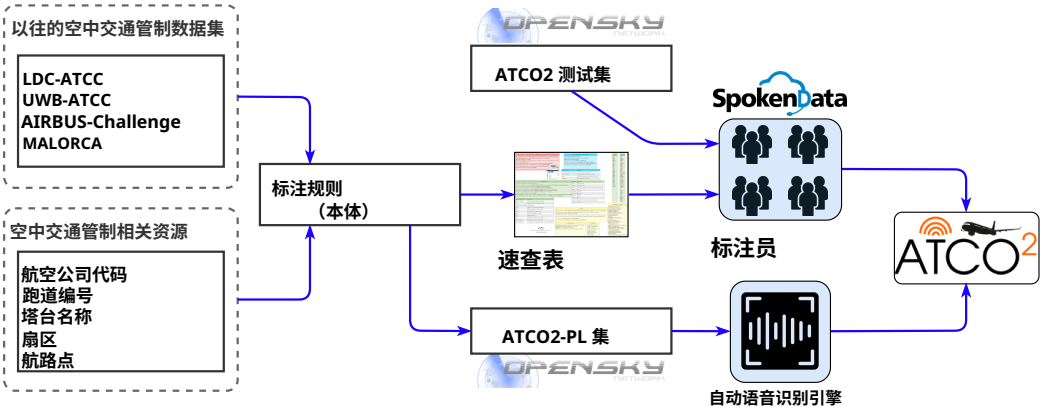


图2. 标注协议工作流程。ATCO2语料库遵循严格的标注协议，该协议基于以往的空中交通管制相关语料库和资源。此外，项目期间还编制了空中交通管制转录文本标注速查表。该速查表或标注内容见附录K。

表2. 在*ATCO2-PL* 集中，经语音活动检测后各机场的语音累计时长（小时）。†英语语言检测（ELD） [0-1] 评分。该评分表示我们的英语语言检测系统在判断空中交通管制通信中是否仅使用英语时的置信度。

国际民用航空组织（机场）-城市										
EETN - Tallinn	EPLB - Lublin	LKPR - Prague	LKTB - Brno	LSGS - Sion	LSZB - Bern	LSZH - Zurich	LZIB - Bratislava	YBBN - Brisbane	YSSY - Sydney	others - others
英语数据（语言评分 ≥ 0.5）†										
131	<1	1762	888	330	699	921	24	170	77	<1
非英语数据（语言评分 < 0.5）†										
2	<1	187	611	83	55	49	26	10	3	<1

这是迄今为止空中交通管制领域中规模最大、内容最丰富的数据集，可用于研究和商业用途。更多信息和详情见 [10]。

- 其次，***ATCO2-test-set-4h***语料库旨在用于英语空中交通管制通信的自动语音识别与理解系统的评估和开发。该数据集由人工标注。如表1所示，该数据集分为两个部分。*ATCO2-test-set-1h*语料库是一个 ∼1小时的开源语料库，可在以下网址免费获取：https://www.atco2.org/data。 *ATCO2-test-set-4h*语料库包含*ATCO2-test-set-1h*语料库，并在此基础上增加了 ∼3 小时的人工标注数据。完整语料库可通过 ELDA 购买，网址如下：http://catalog.elra.info/en-us/repository/browse/ELRA-S0484。

3.2. 数据采集流水线

处理流水线通过遵循配置文件 → worker.py 的 Python脚本实现。配置文件允许我们即时修改处理流程中数据的逻辑和流向。它支持并行化、分叉和条件判断。原则上，worker.py

由全局定义（常量）、模块（局部定义）以及模块之间的链接（有向无环图）组成。处理流水线如图3所示。例如，我们在以往工作中探讨了各项技术的早期实现 [11,19]，例如分段和说话人分离、自动语音识别，或命名实体识别（NER）。所有技术和工具都封装在一组BASH 脚本中，并提供统一的接口。

图3 中的第一行方框指的是分段和解调。首先，天线和录音设备共同采集无线电信号，并将其划分为多个片段，其中包含传输处于“活动”状态的部分；静音部分不会被录制（空中交通管制语音通信使用按键通话）。这一功能由 RTLSDR-Airband 音频录制软件实现，我们通过该软件导出原始 I/Q 信号。接下来，我们使用软件定义无线电工具 CSDR，将这一复数 I/Q 无线电信号转换为波形信号。第一部分在录音设备中完成，第二部分则由 OpenSky Network（OSN）完成。[4]

接下来，我们进行“信噪比（SNR）过滤”（第二行），目的是移除噪声过大的录音。在录音条件较差时，语音可能会变得难以辨认。下一步是“说话人分离”（第三行）。在自动分段的数据中，有些录音包含不止一位说话人。这是个问题，因为我们打算自动转录单个说话人的轮次。此外，对于后续的自然语言处理/口语语言理解（NLP/SLU）任务，区分说话人轮次也很重要。说话人分离通过将音频拆分为仅包含单个说话人的片段，并为其分配说话人标签来解决这一问题。在自动语音识别（ASR）步骤中，我们只需将“语音转换为文本”。这由我们使用 Kaldi 工具包中的工具构建的自动语音识别系统完成 [33]。此步骤的输出是转录文本，其中不可避免地会包含一些错误。为了提高转录文本的准确性，我们会使用上下文信息（监视数据中的呼号列表[5]）。详情请参见第5.2.2 节和 [13,14]。

接下来，转录文本将作为英语检测（ELD）系统的输入。这样做的目的是能够剔除非英语音频数据。典型的先进语言识别系统基于声学建模，并以音频为输入。对于空中交通管制语音，我们无需“识别”非英语语言，因此我们开发了一种“词汇英语检测系统”，以自动语音识别生成的转录文本和置信度评分作为输入。[6] 对于空中交通管制语音，这种方法的效果优于“传统”的声学语言识别方法。最后一项自动操作是“自然语言处理后处理”。目前，处理流程会执行呼号代码提取步骤，并以国际民航组织格式返回属于某架飞机的呼号，例如“DLH77RM”。最后，部分处理后的数据会经过“人工校正”，其余数据则保留自动标签。前者生成了*ATCO2-test-set-4h*语料库，后者则生成了*ATCO2-PL-set*语料库。有关数据收集流程和数据标注的详细说明，请参阅附录I 和附录J。

3.3. 数据标注质量评估

如第 3.2 节末尾所述，采集、处理并自动转录的数据（见图 3）可以由人工进行标注。这样便会生成“金标准标注”，我们利用这些标注来评估所提出的自动语音识别和自然语言处理系统。对 *ATCO2-test-set-4h*语料库 的标注经历了所有这些步骤。然而，在标注之前，我们需要为人工转录选择最易理解且质量更高的数据。由于 OSN 持续记录数据，我们需要选择最易理解且干净的

[4] OpenSky Network 是一个由社区支持的非营利性接收器网络，自 2013 年以来一直持续收集空中交通监视数据。与其他网络不同，OpenSky 会保留完整、未经筛选的原始数据，并向学术界和机构研究人员开放。[5] 呼号列表来自 OpenSky Network 的空中交通监控数据库。[6] 这项工作已在我们此前发表于 Interspeech 的论文中介绍，发表于 2021 [27]。

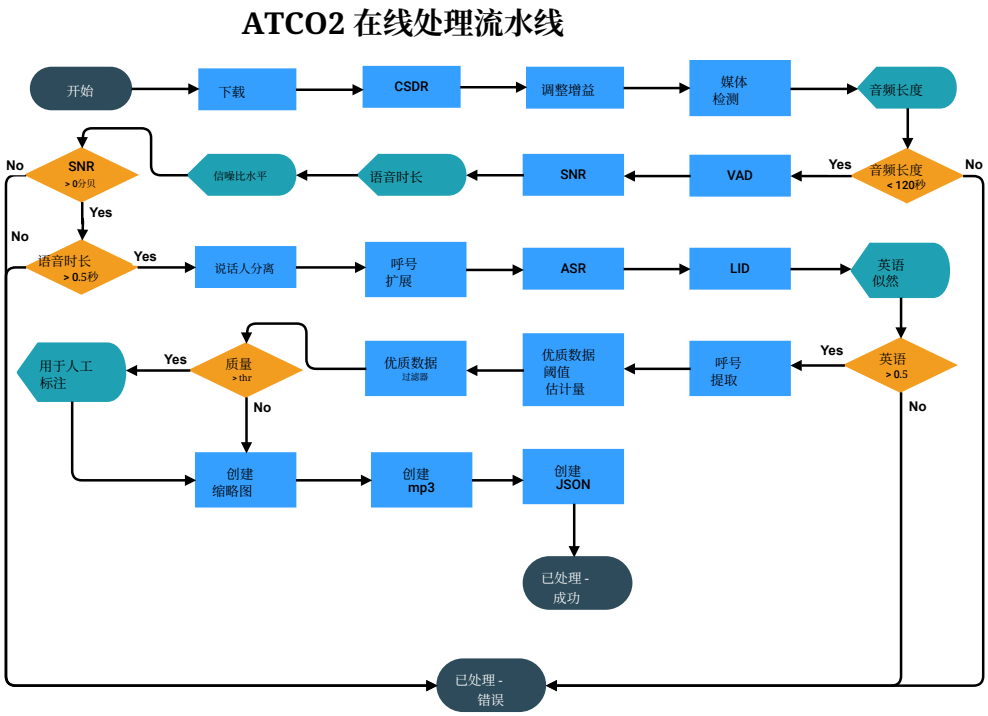


图 3. 用于处理馈线社区收集数据的 *ATCO2*工作流程。首先，数据会发送并存储在 OSN服务器上。音频数据会经过多个模块，以过滤掉噪声水平较高的录音，以及过长或过短的片段。 **蓝色矩形**表示处理步骤。**青色箭头块** 表示内部回调事件，处理流程会通过这些事件向主节点报告进度并发送中间结果。 **橙色菱形** 表示条件，此时会考虑中间结果（例如信噪比水平），以决定是继续处理（音频干净）还是停止处理。处理流程结束时会运行最后一个内部回调。该回调会触发 API 调用 OSN 服务器并传入相应的回调，例如，处理已完成，状态为正常 或 错误。

数据。我们开发了一种评分，用于根据录音质量对录音进行排序。该评分结合了七项指标，用于评估*ATCO2*语料库中每段录音的质量。例如，我们使用公式 1 来衡量、排序并筛选质量最高的空中交通管制通信。随后，这些录音被列入人工标注候选名单（见第 4.2节）。数据标注员生成了真实转录文本和标签，它们构成了*ATCO2-test-set-4h*语料库的一部分。排序评分定义如下：

$$Score = \log(avg_{SNR} + e) + \log(num_{spk} + e) + \log\left(\frac{speech_{len}}{audio_{len}} + e\right) + ELD_{score} \times 3 + avg_{WordConf} \times 3 + \log(wrd_{cnt} + e), \quad (1)$$

其中，

- 平均信噪比 — 提供范围 <0, 40> 内语音的平均信噪比。我们希望信噪比尽可能高，
- 数量说话者 — 提供范围 <1, 10> 内音频中的说话者数量。音频中检测到的说话者越多越好，
- 语音长度 — 表示语音的时长（秒），

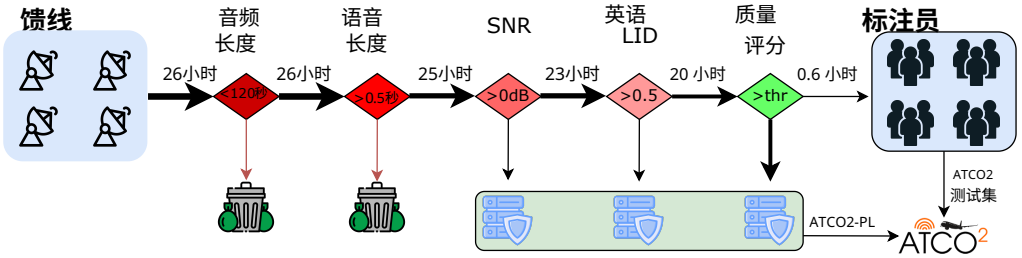


图 4. 从原始数据（数据提供者的录音）到整个处理流程中人工标注的转录文本的数据流产出情况。这是 2022 年 9 月 2 日（DD/MM/YYYY）的一日快照。

- 音频*长度* — 表示音频的总长度。音频中检测到的语音越多越好，
- *ELD*评分 提供音频为英语的“概率”，范围为 <0.0, 1.0>，数值越高越好，

- 平均值*WordConf* - 提供语音识别器 <0.0, 1.0> 的平均置信度。我们需要识别器有信心的数据。数值越高越好，
- 词数 - 提供范围 <0, 150> 内说出的词数。词数越多越好。

图4展示了某一天这些步骤的输出分布。例如，~0从最初的 26 小时音频数据池中，选出 0.6 小时的数据用于金标准标注。我们认为这是一种稳健的质量评分方法，因为它汇集了来自不同系统的信息，例如自动语音识别、信噪比和 ELD 估算。每个步骤输出的每日估算值可在 SpokenData 网站上查看：https://www.spokendata.com/atc。

3.4. 运行特性

我们还测量了处理流水线中各个组件的运行时间。表3列出了各模块所耗时间的相对比例，例如自动语音识别和说话人分离。其他重要环节包括预处理、语音活动检测（VAD）分段和 ELD。预处理包括：获取数据、通过软件无线电解调、分段增益控制、检测媒体格式以及绘制波形。整个处理流水线的一个关键指标是实时因子，即“处理时间”与“音频长度”之比。我们的处理流水线实时因子为 4.47。换句话说，处理过程计算量很大。平均每段五秒长的录音，处理时间为 22 秒。表3展示了“平均”五秒录音中各组件的实际运行时间。

4. 数据收集平台与志愿者社区

本节总结数据收集与分发情况，并简要介绍数据处理过程中各项工作的角色分工。我们还概述了所收集数据的一些总体统计信息。首先，由使用自有接收设备的志愿者采集数据，并将其馈送至 OpenSky Network（见图 5）。这些人通常是航空爱好者，拥有相关运营经验，或对航空技术感兴趣，例如开展相关领域的研究。不过，即使几乎没有航空或技术背景，也可以成为馈线。要成为馈线，必须拥有互联网连接并能使用甚高频接收机。ATCO2语料库论文介绍了一种价格实惠、复杂度低的配置方案 [10]，并提供了配置指南 https://ui.atc.opensky-network.org/set-up。需要注意的是，在某些国家，法律禁止录制空中交通管理（ATM）数据。例如，我们建议

表3. 处理流水线中处理时间的组成部分划分。第二列中的数值对应于一段平均时长为5.016秒的录音。平均值是根据2012年12月4日（DD/MM/YYYY）录制的10334段录音（14.4小时）计算得出的。

处理步骤	时间 [秒]	百分比 [%]
预处理	2,540	11,62
语音活动检测分段	2,436	11,15
信噪比估计	0,655	3,00
说话人分离	7,128	32,62
呼号扩展	0,464	2,12
语音转文本（自动语音识别）	7,021	32,13
英语检测	1,292	5,91
呼号提取	0,069	0,32
后处理	0,245	1,12
总时间	**21,850**	**100,00**

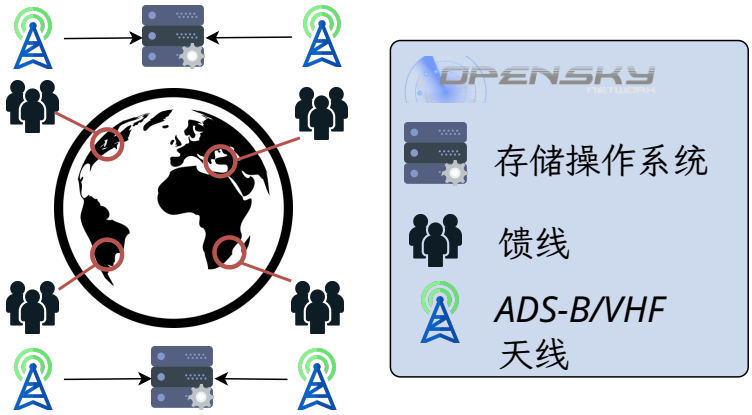


图5. 数据馈送器处理流程。数据用户已设置甚高频接收机，并将数据馈送至 OSN 服务器。

查看**ATC录音收集的法律与隐私方面**一节 [10]。此外，我们也请读者参阅 [34]中有关 ATC 数据收集法律问题的近期研究。

4.1. 平台

高层架构如图 6所示。可以看出，该平台分为三个不同的部分：(i) 馈送器设备、(ii) 后端和 (iii) 前端。ATCO2 在设计阶段确定了这一架构，主要目标是实现整个系统的可扩展性。这意味着要使各部分的复杂度保持相对较低。较低的复杂度有助于：

- 支持与当前 OpenSky ADS-B 系统相近数量的用户；
- 使馈送器设备简单且价格实惠；
- 以简单直观的方式向不同类型的用户提供数据；
- 以简单直观的方式接入外部服务（例如语音标注）；
- 尽可能简化维护和错误处理。

OSN平台的详细概述也列于附录H中。如上所述，该平台分为三个部分。下面我们将逐一介绍平台的各个组成部分。

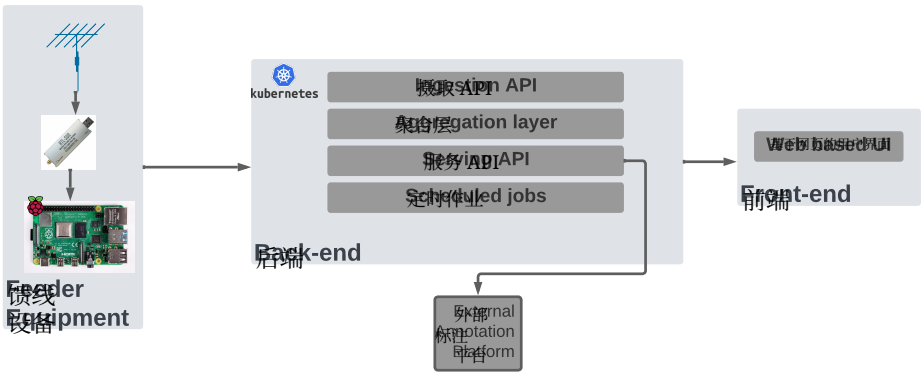


图6. 数据收集平台的高层架构。

馈送器设备：馈送器设备的主要任务是捕获飞行员与空中交通管制员之间的对话，并将数据连同一些相关元数据传送至后端。在录音方面，我们建议将 RTLSDR-Airband 与 RTL-SDR 接收器搭配使用。这是航空爱好者社区中价格实惠且广泛使用的组合，专门用于捕获和传输空中交通管制语音。

馈送软件负责将录音从接收器传输到远程服务器。这是一款相当简单的软件，用于监控 RTLSDR-Airband 的输出目录，并通过 gRPC 连接将找到的任何新数据传输到后端。馈送软件只查找输出文件夹中特定类型的数据，这意味着馈线可以自由选择其他软件来采集和存储语音数据。必须确保输出内容适合馈送软件。我们提供了一份简单的分步指南，以简化设置过程。请访问https://ui.atc.opensky-network.org/set-up。

后端： 后端的主要任务是：(i) 存储录音、转录文本及其他相关元数据；(ii) 为外部用户提供接口。此处的外部用户包括数据馈送器、转录服务提供商、数据用户，以及其他任何参与数据集建设或使用数据集的各方。后端部署在 Kubernetes（一种开源容器编排系统）上。从图6可以看出，其中涉及多个处理层。这些层包括：

- 摄取 API：接收录音片段和元数据，并将其排入 Kafka/兼容 S3 的对象存储中的队列，以供处理。
- 聚合层：将原始数据转换为 FLAC 音频，存储元数据，并使用 Kafka Streams、S3 和服务 API 触发转录。
- 服务 API：提供外部接口，用于使用元数据、进行存储，以及获取转录文本和统计数据。
- 定时作业：运行不属于流处理过程的任务，例如统计数据聚合和数据清理。

后端通过 API 接入，相关文档详见 https://api.atc.opensky-network.org/q/swagger-ui。若要访问后端并使用可用的 API，用户需要在 https://auth.opensky-network.org/auth/ 注册，联系 OpenSky Network（mailto: `contact@opensky-network.org`），并简要说明访问用途。

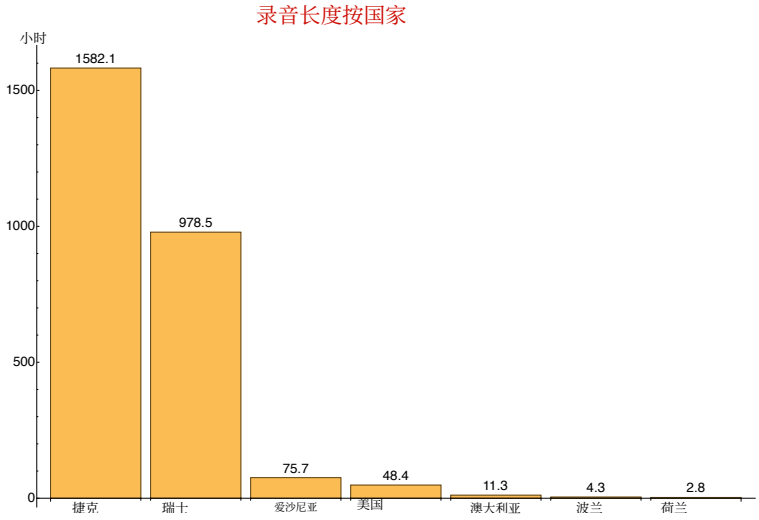


图 7. 从服务开始至 *2023* 年 *3* 月 *5* 日每个国家的录音长度。给出了录音长度超过 1 小时的国家。

前端： 前端是一个网页（https://ui.atc.opensky-network.org/），可用于查看公开统计数据、文档链接（例如 API 文档）和外部网页（例如 SpokenData 转录服务）。此外，用户还可以在此设置接收器、查看接收器性能统计数据等。

4.1.1. 统计

自服务于2023年3月5日开通以来，ATCO2项目已录制了来自14个不同国家的24座机场的语音。图7 和图8展示了各国和机场名称，以及相应的录音时长。请注意，仅纳入再研磨时长为 ≥ 1 小时的地点。

4.2. 数据标注员

除了数据提供者，还有另一类志愿者为项目作出了贡献，并将在未来继续贡献。他们被称为“标注员”。数据标注员是负责记录空中交通管制语音通信转录文本的志愿者，其中包括标注说话人和命名实体（例如呼号和指令）。ATCO2项目既依靠志愿者，也依靠付费转录员。如图3所示，我们的数据处理流程会为每段通信生成转录文本和自然语言处理标签。借助人工智能工具进行预转录，我们能够加快整体转录过程。[7] 人工转录 [8] 数据量为一个四小时测试集，即 *ATCO2-test-set-4h*语料库。如图4所示，数据标注员是参与人工数据标注步骤的最后一环。

5. 技术

本节介绍基于 ATCO2语料库开发的主要工具，并列出一些可借助该语料库探索的潜在主题。此外，请注意，ATCO2语料库并

[7] 如果您有兴趣成为标注员，请在 SpokenData转录平台上创建账户：http://www.spokendata.com/atco2。[8] 本语料库是 ELDA 目录中发布的数据的一部分：http://catalog.elra.info/en-us/repository/browse/ELRA-S0484/

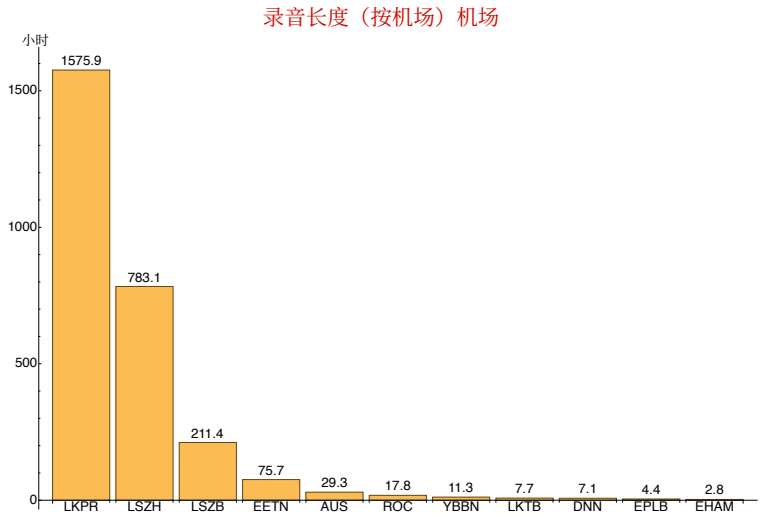


图8. 从服务开始至*2023*年*3*月*5*日各机场的录音时长。此处列出了录音时长超过1小时的机场。

不局限于本文涉及的领域，例如自动语音识别或自然语言处理。我们希望相关领域的研究者能够基于该语料库开展工作，推动面向空中交通管制的语音和文本技术发展。

5.1. 自动语音识别

ATCO2项目的主要组成部分之一是强大的自动语音识别系统，旨在为所收集的空中交通管制数据提供高质量的自动转录。自动语音识别系统经过训练，可为输入声学信号预测最佳文本翻译。形式上，自动语音识别旨在给定带噪声的声学观测序列时，从某种语言中所有可能的词语组合（或句子）构成的集合中，找到概率最高的词语输出序列候选。端到端自动语音识别模型学习从语音到输出文本的直接映射：

$$y^* = \underset{y}{\text{argmax}} \log P(y|x) \tag{2}$$

混合型（传统）自动语音识别系统结合了三个分别训练的模型：声学模型（AM）、发音模型和语言模型（LM）。该模型计算条件概率 $p(W|S)$，其中 W 是词语序列（$W = \{w_1,\ldots\ldots, w_n\}$），$S$ 是表示声学观测的输入特征向量序列（$S = \{s_1,\ldots\ldots, s_t\}$），而$\mathcal{V}$ 是词汇表 [35,36]（公式5）：

$$\hat{W} = \underset{W\in\mathcal{V}}{\text{argmax}}\, p(W|S) \tag{3}$$

$$= \underset{W\in\mathcal{V}}{\text{argmax}}\, p(S|W)p(W) \tag{4}$$

$$= \underset{W\in\mathcal{V}}{\text{argmax}} \sum_P p(S|P)p(P|W)p(W) \tag{5}$$

其中，$p(S|P)$ 是一个 AM，$p(P|W)$ 是一个发音模型，而$p(W)$ 是一个 LM。传统处理流程模型的优势之一是能够更透明地优化目标函数 [37]。此外，LM 使用未配对的文本数据进行训练，并且可以轻松适配特定领域。这使得传统模型具有更高的灵活性，因此便于应用于空中交通管制等工业项目。

5.1.1. 训练数据和测试数据

训练数据配置： 为了衡量使用自动转录数据（ATCO2-PL 数据集）与使用完全监督的金标准标注的效果，我们定义了三种训练场景。

• 情景 ***a***）仅使用监督数据：我们采用公开和私有空中交通管制数据库的混合数据（带有金标准标注的录音）。其中包含 ∼190 小时的音频数据（速度扰动后为 573 小时）。

• 场景 ***b***）仅使用 ***ATCO2-PL*** 语料库中的 ***500*** 小时数据集：我们仅使用 ATCO2-PL 语料库中的 500 小时子集（参见介绍性论文 [10]）。
• 场景 ***c***）仅使用 ***ATCO2-PL 2500*** 小时数据集：与 **场景 b**）相同，但我们使用的不是 500 小时子集，而是多 5 倍，即 2500 小时子集。该子集仅用于训练一个基于混合模型的自动语音识别模型（CNN-TDNNF），以测试第 5.2.2 节中的增强实验。

测试数据配置： 如表 4所示，自动语音识别评估使用了两个 ATCO2 测试集：ATCO2-test-set-1h（简称 ATCO2-1h）和 ATCO2-test-set-4h（简称 ATCO2-4h）。第 5.2.2 节所述的增强实验也使用了相同的测试集。

5.2. 传统自动语音识别

为了从 ATCO2 语料库音频中获得质量尽可能高的自动转录文本，我们使用了一个仅基于空中交通管制数据训练的强混合模型。针对上述每种场景，我们都训练了一个基于混合架构的模型。对于 **场景 a)**，我们构建了一个声学模型，纳入了所有可用的 190 小时数据；经过语音增强后，数据总量为 573 小时。模型词典包含 30’832 个单词。我们使用 Kaldi 工具包 [33]训练声学模型。该系统遵循标准的 Kaldi 配方，使用梅尔频率倒谱系数和 i 向量特征。标准链式训练基于无格最大互信息（LF-MMI [38]，其中包括 3 倍速度扰动和三分之一帧子采样。声学模型为 CNN-TDNNF。语言模型是一个三元语法模型，使用与声学模型相同的数据训练，并额外加入来自公共资源的文本数据，例如航空公司名称、机场、国际民航组织字母表以及欧洲的航路点。

结果与分析： 结果如表4所示。我们比较了使用相同的传统 CNN-TDNNF架构、但基于不同数据训练的三个模型：场景 a)、b) 和 c)（见 5.1.1）。表 4 中的模型 a) 使用针对 ATCO2 的“域外”监督数据进行训练。模型 b) 和 c) 使用“域内”的 ATCO2 数据进行训练，二者唯一的区别在于训练集大小：500 小时与 2500 小时。可以看出，使用完全无监督的数据进行训练也能取得良好的性能。然而，将无监督数据的规模从 500 小时增加到 2500 小时，并不会带来太大改善：对于 ATCO2-1h 和 ATCO2-4h，词错误率（WER）分别仅从 18.1% 降至 17.9%，以及从 25.1% 降至 24.9%。

5.2.1. 端到端自动语音识别

与基于混合模型的自动语音识别不同，自动语音识别还有另一种范式，即端到端自动语音识别。在这种范式中，我们旨在直接将语音转录为文本，而无需建立输入特征与输出词语或字符之间的对齐关系（也就是基于混合模型的自动语音识别中的标准流程），见公式5。近期关于编码器-解码器自动语音识别的研究表明，这一步骤可以省去。端到端自动语音识别可分为两类：连接时序分类 [39] 和基于注意力的编码器-解码器建模 [40]。有关编码器-解码器模型的其他研究，见 [41–45]。端到端自动语音识别旨在减少对专家知识的需求。这使整个自动语音识别开发流程更加简单，因此可能对空中交通管制产生重大影响 [46]。正如本研究所做的那样，我们首先关注数据的新颖性及其收集和准备工作，并不比较不同的自动语音识别架构。因此，端到端自动语音识别的结果留待其他论文讨论。

表 4. 使用不同数据（场景 (a-c)）训练的 CNN-TDNNf 模型在公开 ATCO2 测试集上的词错误率结果。

模型	测试集	
CNN-TDNNF	**ATCO2-1小时**	**ATCO2-4小时**
场景 a）仅使用 573 小时的监督式数据集	24.5	32.5
场景 b）仅 ATCO2-PL 500 小时数据集	18.1	25.1
场景 c）仅使用 ATCO2-PL 2500 小时数据集	17.9	24.9

5.2.2. 呼号增强

要进一步改善自动语音识别系统的预测结果，除了语音输入，还可以利用上下文中可获得的其他信息。对于空中交通管制领域，这类上下文信息可能是从雷达接收到的数据。雷达持续记录当前处于空域中的飞机，并列出这些飞机各自的唯一标识符，即“呼号”。借助雷达数据，我们可以准确了解哪些呼号特别可能出现在对话中。这些信息使我们能够让系统输出偏向已登记的呼号，从而提高正确识别它们的概率。呼号通常由航空公司名称、字母和数字组成；在语音中，它会转化为一串词语。在自动语音识别中，可以在加权有限状态转换器解码过程中，通过调整预测图（称为“格结构”）中的权重，来提升目标词语序列的优先级 [5,47–49]。格结构的重新评分通过有限状态转换器操作中的组合来完成：将自动语音识别系统生成的格结构与根据目标词语序列和折扣权重创建的有限状态转换器组合（公式6）：

$$biased_Lattices = Lattices \circ biasing_FST \tag{6}$$

将格状图偏向上下文中的呼号，通常可以显著提升最终输出中对这些呼号的识别效果（表 5 ）。针对空中交通管制数据开展的不同实验结果证明，在自动语音识别预测的基础上应用格状图重新评分方法，能够提高自动转录的准确率，首先是呼号方面 [5,18]。因此，所有 ATCO2 数据的转录均采用了格状图重新评分。

结果与分析： 表 5中报告了域外（由空中交通管制数据监督训练）和域内（ATCO2-500h/2500hr）空中交通管制模型的结果。两个声学模型均按照标准 Kaldi 配方，采用 CNN-TDNNF架构进行训练，具体如 5.2 所述。

与仅对真实标注（GT）呼号进行偏置相比，偏置n元语法可用于现实场景和实时自动语音识别：每当有新数据从雷达传来，系统便会即时生成新的上下文有限状态转换器。此处给出了对真实标注呼号进行偏置的结果，以说明该偏置方法的预言机性能。总体而言，使用n元语法偏置进行解码总能提升性能，尤其是呼号识别方面：在 ATCO2-test-set-1h 和 ATCO2-test-set-4h 测试集上，呼号识别的相对提升分别为 15.0% 和 12.8%，整个话语的相对提升分别为 3.4% 和 2.4%。

偏置有限状态转换器的大小取决于要提升权重的实体数量。实体过多可能会降低偏置方法的有效性，因为被提升权重的非真实实体越多，正确序列的突出程度就越低。前面的结果表明，偏置有限状态转换器的最佳大小在很大程度上取决于数据，但通常当上下文实体的数量超过 1000 [50]时，性能便开始下降。在我们的实验中，ATCO2-test-set-4h中每条话语平均包含214个偏置实体，而ATCO2-test-set-1h语料库中每条话语平均包含140个偏置实体。

表5。 ATCO2语料库增强实验的结果。结果列出了使用全部监督数据、500小时或2500小时ATCO2语料库训练的CNN-TDNNf模型的结果。每个区块中的最佳结果以**粗体突出显示**。每列中的最佳结果以下划线标出。†1小时公开测试集。
‡4小时完整测试集。结果通过离线CPU解码获得。¶词错误率仅针对话语中构成呼号的词序列计算。

增强	ATCO2-test-set-1h†			ATCO2-test-set-4h‡		
	词错误率	实体词错误率¶	准确率	词错误率	实体词错误率¶	准确率
	场景 a）仅使用监督数据集					
基线	24.5	26.9	61.3	32.5	36.7	42.4
一元语法	24.4	25.5	63.2	33.1	35.0	45.8
N元语法	23.8	23.8	66.4	31.3	33.7	47.9
GT增强	**22.9**	**19.1**	**75.2**	**29.7**	**29.1**	**58.5**
	场景 b）仅使用 ATCO2-PL 500 小时数据集					
基线	18.1	16.2	71.2	25.1	24.8	62.6
一元语法	19.1	14.6	74.2	26.0	22.8	65.6
N元语法	17.5	13.5	75.3	24.3	21.4	66.6
GT增强	**16.3**	**6.9**	**88.9**	**22.5**	**13.0**	**82.9**
	场景 c）仅使用 ATCO2-PL 2500 小时数据集					
基线	17.9	16.7	70.5	24.9	24.2	62.0
一元语法	18.3	14.4	73.8	25.6	22.0	65.9
N元语法	17.3	14.2	74.3	24.3	21.1	66.5
GT增强	**15.9**	**6.5**	**89.4**	**22.2**	**12.5**	**83.9**

5.3. 空中交通管制通信的自然语言理解

自然语言理解（NLU）是自然语言处理（NLP）的一个子领域，关注理解和解释人类语言的能力。自然语言理解涉及开发能够从文本和/或口语交流中提取含义和意图的算法与模型。自然语言理解包含多个子任务，包括：（i）命名实体识别 [51,52], ，旨在识别文本中的实体，例如人物、地点和组织。（ii）词性标注（POS），用于识别句子中每个词的语法角色 [53], ，类似于序列分类（见第5.3.2节）。（iii）情感分析，用于识别一段文本的情感色调 [54]。（iv）关系提取，用于识别文本中实体之间的关系。（v）问答，用于理解和回答自然语言问题。以下小节将介绍使用ATCO2语料库可以开发的各个自然语言理解子模块，例如图9中所示的模块。

5.3.1. 空中交通管制通信的命名实体识别

在空中交通管制通信中，自然语言理解可用于自动分析和理解飞行员与空中交通管制员之间口头信息的含义。这有助于提高通信的准确性和效率，也有助于识别紧急情况及其他关键事件。此外，自然语言理解还能提取重要信息，例如航班号、呼号或机场代码，从而帮助空中交通管制员更高效地管理空中交通。总的来说，在空中交通管制中运用自然语言理解有助于提高通信的准确性和效率、增强安全性，并为分析和决策提供有价值的数据。

在这项工作中，主要任务之一是理解并提取空中交通管制通信中的高层信息。因此，我们开发了一个用于提取这些信息的NER系统

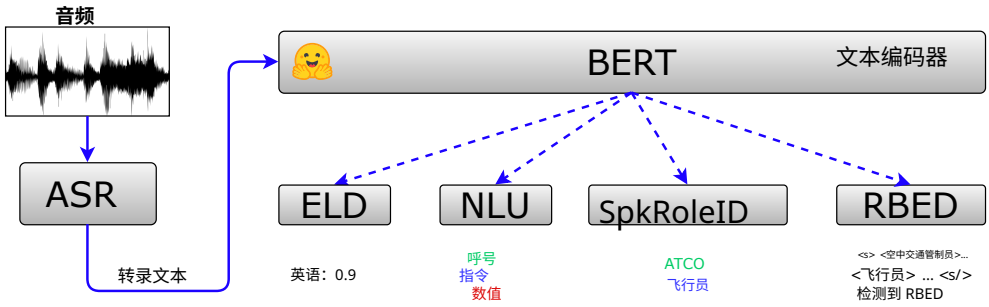


图9. 利用 *ATCO2*语料库可实现的主要自动语音识别与理解任务。ELD：英语检测；NLU：自然语言理解，例如呼号标注；SPKRoleID：说话人角色识别；RBED：复诵错误检测。

，如图 10 （a）所示。例如，请看以下转录的通信内容（摘自图 1）：

自动语音识别转录文本： 34号左跑道准许着陆，中国南方航空325，

将通过 NER系统转换为高级实体格式：

输出： <数值> 跑道 三 四 左 </数值> <指令> .....准许.....着陆 </指令><呼号> 中国南方航空325 </呼号>。

在本研究中，我们开发了两个基于 Transformer 的系统 [55]，用于从空中交通管制通信中提取并标注这些信息，即一个预训练的 BERT [56] 模型和一个 RoBERTa [57]模型。

实验设置： 我们针对命名实体识别任务对预训练的 BERT 和 RoBERTa[9] 模型进行微调（见图9）。我们从 HuggingFace [58,59]下载预训练模型。训练时，我们使用完整的 ATCO2-test-set-4h，其中包含 ∼3k 个句子。在此数据集中，每个词都与一个预定义类别一同标注，类别如下：呼号、.....指令、 数值以及UNK（其他所有内容）。为微调模型，我们在 BERT 模型之上添加一个维度为 8 的前馈网络层。[10]由于缺少金标准标注，我们采用五折交叉验证方案来避免过拟合。有意开发自己的空中交通管制命名实体识别系统的读者，可访问 ATCO2 语料库的开源 GitHub 仓库。[11] 我们在 NVIDIA GeForce RTX 3090 上对每个模型进行微调，共执行 ∼10k 步。实验期间，我们使用线性学习率调度器，初始学习率为 $\gamma = 5e{-}5$，Dropout [60] 率为$dp = 0.1$，并采用 GELU 激活函数 [61]。我们还使用梯度范数裁剪 [62] 进行正则化，并使用 AdamW 作为优化器 [63]。交叉验证方案中的每个模型均采用 32 的有效批次大小。

评估指标： 我们使用名为 F-score 的二元分类指标，对 BERT 和 RoBERTa NER 系统进行评估。具体而言，F1分数在公式 9 中定义，表示精确率和召回率的调和平均值。召回率的定义见公式 8，其计算方式为 *TP* 占所有应被识别为正类的样本（包括假阴性（*FN*））的比例。

[9] 预训练版本的 bert-base-uncased [56]，包含 1.1 亿个参数，网址：https://huggingface.co/bert-base-uncased。预训练版本的 `RoBERTa-base` [57] 包含 1.23 亿个参数，网址：https://huggingface.co/roberta-base[10] 我们为每个类别定义了两个输出。请参阅 [12] 第 3.3 节中的类别结构，以及 [5]。[11] GitHub 仓库：https://github.com/idiap/atco2-corpus

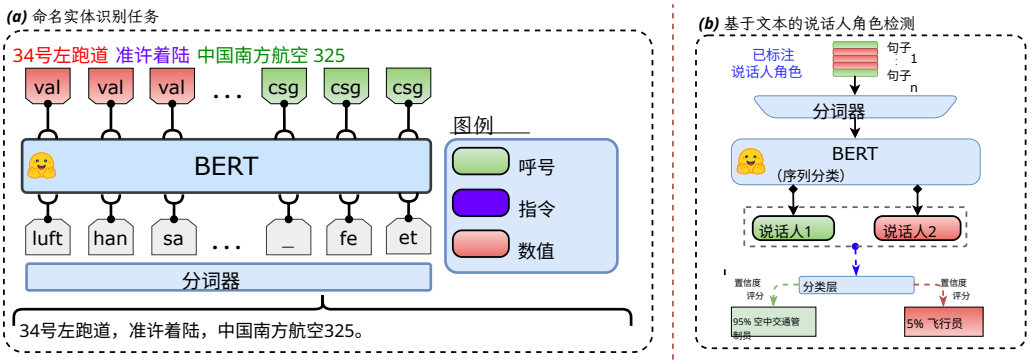


图10。 （*a*）基于序列分类（*SC*）的命名实体识别和（*b*）说话人角色检测模块。两个系统均基于在空中交通管制数据上对预训练BERT [56] 模型进行微调。NER系统识别呼号、指令和数值，而SC则为输入序列分配说话人角色。摘自 [10]。

精确率，如公式 7所述，是真正例（*TP*）结果占所有正例结果（包括假正例（*FP*））的比例：

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{9}$$

结果与分析： NER系统的性能在ATCO2 测试集-4h语料库上采用五折交叉验证方案进行评估，并使用不同的训练随机种子进行了五次微调运行。表6 展示了两个基于Transformer的 [55] 模型，即BERT-base和RoBERTa-base，在呼号、指令和数值类别上的性能指标。尽管两个模型取得了相近的F1分数，我们仍对基于BERT的命名实体识别系统进行分析；该系统在 呼号 类别上取得了超过97%的F1分数，而.....指令 和 ~~~数值类别则相对落后，F1分数分别为81.9%和87.1%。我们推测，与另外两个类别——数值和呼号——相比，指令类别包含更高的复杂性。数值主要由定义明确的关键词（例如飞行高度层）以及其后的基数词（例如“一百”）组成，而呼号则遵循由航空公司代码和一组按照国际民航组织格式读出的数字或字母构成的明确定义结构 [28]。这些特征使NER系统更容易正确检测它们。

一种提升指令和数值类别命名实体识别系统性能的潜在方法，是根据监视数据的具体情况，实时纳入合理的指令和数值。可使用第5.2.2节所述的提升技术来实现。尽管表5报告了增强呼号的结果，但仍需进一步研究，以评估增强对指令和数值类别的影响。

5.3.2. 基于文本的说话人角色检测

序列分类（SC）是一种机器学习任务，涉及为一组数据点序列 [64,65] 分配一个标签或类别。序列中的数据点可以有多种类型，例如文本、音频或数值数据，而分配给该序列的标签

表 6. 命名实体识别（NER）系统中呼号、指令和数值类别的不同性能指标。为减轻过拟合，结果是在*ATCO2-test-set-4h*语料库上采用五折交叉验证方案进行平均得到的。我们使用不同的训练随机种子（2222/3333/4444/5555/6666）进行了五次微调。结果基于两个基于 Transformer 的模型报告。@P、@R 和 @F1 分别表示精确率、召回率和 F1分数。

模型	呼号			指令			数值		
	@P	@R	@F1	@P	@R	@F1	@P	@R	@F1
BERT-base	97.1	97.8	97.5	80.4	83.6	82.0	86.3	88.1	87.2
RoBERTa-base	97.1	97.7	97.5	80.2	83.7	81.9	85.6	88.6	87.1

也可以有不同的类型，例如二元（如正面或负面情感 [54]）或多类别。序列分类还可用于自动将空中交通管制通信序列分类到不同类别中。这项技术既适用于音频，也适用于文本数据，因此是一种用途广泛的工具，可用于对当前通信进行高层次的理解。

在无法获取按键通话信号、且仅存在单声道通信信道的场景中，辨别说话人身份可能颇具挑战。因此，区分目标通信中的空中交通管制员和飞行员至关重要。作为一种潜在的解决方案，我们提出了一种基于说话人分类（SC）的说话人角色检测（SRD）系统。该系统以文本为输入，并输出通信所属的类别，即由空中交通管制员或飞行员发出。近年来，人们越来越关注使用深度学习技术（例如基于Transformer 的模型） [55], 来提升空中交通管制通信中用于说话人角色检测的说话人分类性能。本文对三种此类模型进行了消融研究：(i) BERT [56], (ii) RoBERTa [57], 和 (iii) DEBERTA [66]。研究表明，这些模型在各种序列分类任务（包括空中交通管制中的说话人角色检测）上均达到了最先进的性能。所提出的说话人角色检测系统如图 10(b) 所示。

总体而言，SRD 和说话人分离（见第 5.3.3节）任务可以利用空中交通管制对话遵循定义明确的词典和具有简单语法这一特点。这种标准通话用语由国际民用航空组织制定 [28] ，旨在保障安全并减少空中交通管制员与飞行员之间的通信误解。因此，先前的研究已展示了在文本级以端到端方式执行 SRD 的潜力，本文也采用了这种方法，详见 [12,67,68]。

实验设置： SRD系统基于预训练模型构建（BERT [56],RoBERTa [57], 和 DEBERTA [66]），这些模型从 HuggingFace [58,59]下载。此处的实验设置与 NER系统所述完全相同，包括训练超参数。详情请参阅第 5.3.1节。不过，SRD模型是在说话人分类任务上进行微调，而非命名实体识别任务。此外，我们定义了一个包含2个单元（类别）的输出层，分别对应空中交通管制员和飞行员。

结果与分析： 我们在 ATCO2-test-set-4h语料库上评估 SRD 系统。与 NER 系统不同，这里我们可以使用两个训练语料库：LDC-ATCC[12] 和 UWB-ATCC[13] 数据集。我们从两个方面评估 SRD：（i）针对空中交通管制通信的不同预训练模型的消融实验；

[12] 空中交通管制语料库（LDC-ATCC）语料库，网址见：https://catalog.ldc.upenn.edu/LDC94S14A。该语料库包含空中交通管制通信领域的自动语音识别音频录音。我们结合元数据和转录文本，开展空中交通管制自然语言理解研究，即说话人角色检测。数据文件的采样率为 8 kHz，采用 16 位线性编码，连续监测，且未使用静噪或静音消除。[13] 西波希米亚大学的 UWB-ATCC 语料库可通过以下网址免费下载：https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-CCA1-0。UWB-ATCC 包含

表 7. 空中交通管制员/飞行员 ATCO2-test-set-4 测试集上说话人角色识别的 F1分数。使用三种不同的基于 Transformer 的模型（BERT [56], RoBERTa [57], deBERTa-V3 [66]）。所有模型均为"base"版本，例如 `bert-base`。 **粗体** 数字表示各数据划分上的最佳表现，即 **空中交通管制员** 或 **飞行员**。结果采用五折交叉验证方案，在*ATCO2-test-set-4h*语料库上取平均值，以减轻过拟合。每轮微调均使用不同的训练随机种子（2222/3333/4444/5555/6666）运行五次。

训练语料库	BERT		DEBERTA		ROBERTA	
	空中交通管制员	飞行员	空中交通管制员	飞行员	空中交通管制员	飞行员
LDC-ATCC	82.4	79.2	82.4	79.6	84.0	80.2
UWB-ATCC	86.2	83.2	86.8	84.0	87.0	82.8
↪ +LDC-ATCC	**87.6**	**85.2**	**88.8**	**85.8**	**88.0**	**84.2**

（ii）低资源和增量训练场景。

(i) 预训练模型和训练数据类型的影响分析 在此情境下，我们评估预训练模型和训练数据对空中交通管制数据说话人角色检测任务的影响。为此，我们比较了三个基于Transformer的模型的表现， [55]包括 BERT、RoBERTa 和 deBERTa-V3。这些模型分别使用两个不同的语料库 LDC-ATCC 和 UWB-ATCC 进行训练，并在 ATCO2-test-set-4h语料库上进行评估。表中分别列出了空中交通管制员和飞行员说话人的说话人角色检测 F1分数。7。结果显示，所有模型的 F1分数相近，空中交通管制员的得分为 87% 至 88%，飞行员的得分为 84% 至 85%。这些发现表明，空中交通管制数据的说话人角色检测任务对预训练模型的选择并不十分敏感。不过，我们观察到，使用 UWB-ATCC 训练的模型优于使用 LDC-ATCC 训练的模型，F1分数最高提升了 4 个百分点。例如，使用 ***LDC-ATCC*** 训练的 ***BERT*** 模型与使用 UWB-ATCC 训练的模型相比， → 空中交通管制员的得分为 82.4% → 86.2%，飞行员的得分为 79.2% → 83.2%。此外，我们发现，合并两个数据集可使 F1分数提高 1 个百分点。总的来说，本研究强调了为面向空中交通管制数据的说话人角色检测任务选择合适训练数据的重要性，并表明使用多个数据集能够提升性能。研究结果还表明，预训练模型的选择对空中交通管制数据的说话人角色检测任务影响相对较小。

(ii) 数据量对说话人角色检测的影响分析 在本研究中，我们旨在评估文本样本数量对说话人角色检测性能的影响。该分析结果如图 11左侧面板所示，其中绘制了 ATCO2 测试集上的 F1分数与样本数量的关系，横轴采用对数尺度。有趣的是，我们发现，仅需 100 个样本，即可在说话人角色检测任务中达到相当不错的 60% F1分数。值得注意的是，UWB-ATCC 对 BERT 模型似乎更有信息量，仅使用 100 个训练样本便取得了 71% 的 F1分数。将训练数据增至 1000 个样本后，性能进一步提升，F1分数接近 80%（LDC-ATCC + UWB-ATCC）。鉴于空中交通管制通信的标注通常成本高昂且耗时，这些发现意义重大。图 11右侧面板展示了箱线图，显示 BERT 模型在使用不同训练随机种子对说话人角色检测进行微调时的性能变化。每个箱体表示在五折交叉验证方案下，模型在空中交通管制员子集与飞行员子集之间的变化。总体而言，结果表明，增加训练数据可以提升性能，并带来更

空中交通管制通信录音。语音经过人工转录，并标注了说话人信息，因此可用于说话人角色检测。

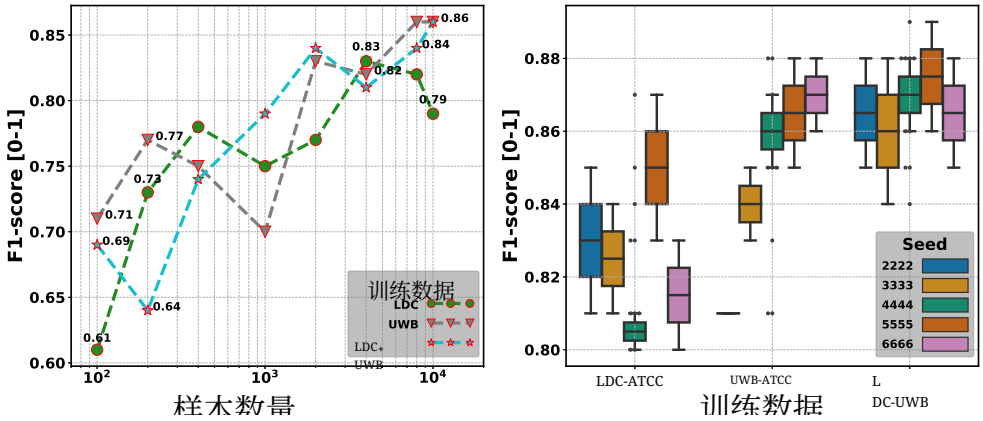


图11. 说话人角色检测系统的指标（见 *[10]*）。指标仅在 *ATCO2- test-set-1h* 语料库 上报告，使用 `bert-base-uncased`模型，并采用表 1 中不同的数据集进行训练。左图：消融实验中F1分数与用于训练系统的样本数量之间的关系。右图：使用不同训练随机种子训练的模型的F1分数。箱线图展示了将测试集划分为ATCo子集和飞行员子集时的性能变化。

结果一致。这些观察结果凸显了为说话人角色检测任务选择合适训练集规模的重要性。

5.3.3. 基于文本的说话人日志

除了仅检测给定空中交通管制通信中的角色（例如，SRD）之外，有时同一录音/通信中会出现多个片段。解决这一问题的任务称为说话人分离（SD）。SD回答的问题是 “谁在什么时候说话？”。在此，系统接收音频信号或录音（在我们的案例中则是文本），并检测说话人变更或进行分段，同时识别说话人角色。SD系统的主要组成部分包括：(i) 分段，(ii) 嵌入提取，(iii) 聚类，以及 (iv) 标注（与SRD类似）。SD通常在声学层面进行；此前基于梅尔滤波器组斜率和线性滤波器组斜率的研究见 [69]。关于x向量等说话人判别嵌入的研究见 [70]，，近期则在 [71,72]，中研究了变分贝叶斯隐马尔可夫模型（VBx）；该模型是ATCO2数据收集阶段所用的SD系统，见第 3.2 节。最先进的SD系统基于端到端范式，称为端到端神经说话人日志（EEND）。这种方法由 [73,74] 提出。在该方法中，SD模型经过联合训练，以执行提取和聚类。与SRD不同，这里我们只使用了BERT [56] 预训练模型。

实验设置： SD 系统基于从 HuggingFace 下载的预训练 BERT 模型构建 [58,59]。与 NER 和 SRD 系统一样，此处的实验设置相同，其中也包括训练超参数。[14] 详情请参阅第 5.3.1 节。SD 模型在 NER 任务上进行微调，其中每种说话人角色（ATCo 或飞行员）对应一个类别。因此，每个类别有 2 个标签，总共 4 个类别。有关基于文本的 SD 的更多信息，请参阅我们发表于 2022年IEEE语音语言技术研讨会（SLT 2022）的论文，见 [12]。

评估指标： 为了评估基于文本的说话人日志（SD）系统，我们使用 Jaccard错误率（JER）指标。JER 是一种较新的指标，在 [75] 中提出，与说话人分离相对应。JER 旨在

[14] 查看 ATCO2 的开源 GitHub 存储库： https://github.com/idiap/atco2-corpus

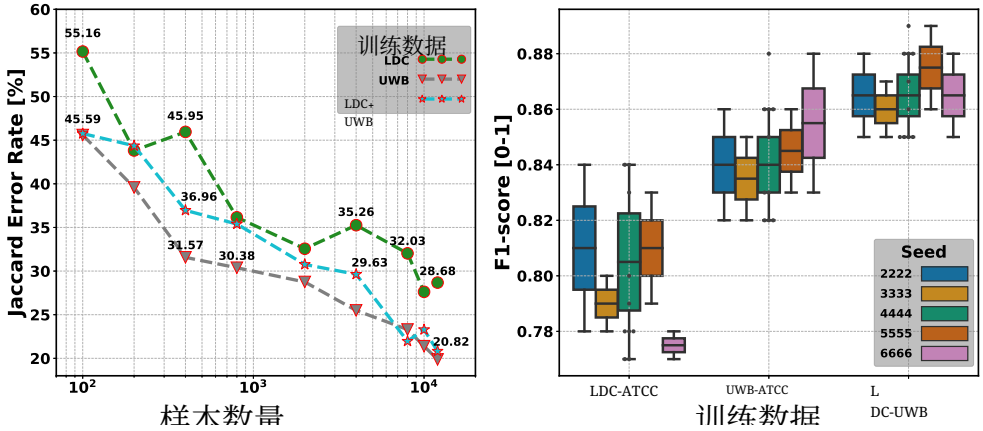


图12. 基于文本的说话人分离系统的指标（见 *[10,12]*）。指标仅在*ATCO2-test-set-4h*语料库上报告，采用`bert-base-uncased`模型，并使用表1中的不同数据集进行训练。左图：消融实验中，Jaccard错误率随用于训练系统的样本数变化的情况。右图：使用不同训练随机种子训练的模型的F1分数。箱线图展示了按空中交通管制员和飞行员子集划分测试集时的性能变化。

避免占主导地位的说话人可能造成的偏差。也就是说，JER评估会平等地评估所有说话人。JER 的定义见公式10：

$$JER = 1 - \frac{1}{\#\text{speakers}} \sum_{\text{speaker}} \max_{\text{cluster}} \frac{|\text{speaker} \cap \text{cluster}|}{|\text{speaker} \cup \text{cluster}|}, \tag{10}$$

其中，(i) speaker 是从参考语音中选出的说话人，(ii) $max_{cluster}$ 是系统中与当前选定说话人的重叠时长最大的聚类。

结果与分析：我们在 ATCO2-test-set-4h语料库上评估 SRD 系统。与 NER 系统不同，但与 SRD 类似，我们可以使用两个训练语料库：LDC-ATCC 和 UWB-ATCC 数据集。我们在以下一种设置下评估 SD：(i) 低资源和增量训练场景。

(i) 数据量对基于文本的说话人分离的影响分析 本研究旨在评估文本样本数量对说话人日志性能的影响。分析结果如图12左图所示，其中绘出了 ATCO2-test-set-4h 上的 JER（越低越好）与样本数量的关系，横轴采用对数尺度。我们发现，只需 100 个样本即可达到 45.6% 的 JER 评分（LDC-UWB）。与 SRD 类似，UWB-ATCC 数据集似乎能为说话人日志系统提供更多信息。例如，在 1000 个样本的情况下，使用 UWB-ATCC 可使 JER 绝对值降低 5%。此外，将训练数据增加至 1 万个样本后，性能有所提升，JER 评分接近 20%（LDC+UWB）。关于空中交通管制通信中基于文本和基于声学的说话人日志更恰当的比较，可参见我们此前的研究 [12]。此外，图 11 右图给出了箱线图，展示了基于 BERT 的说话人分离模型在使用不同训练随机种子进行微调时的性能变化。每个箱体表示在五折交叉验证方案中，模型对提出的两类说话人（空中交通管制员和飞行员）的分类表现变化。结果以 F1 分数列出。总的来说，与 LDC-ATCC 数据集相比，UWB-ATCC 数据集能为说话人日志模型提供更多信息。

5.3.4. 通信结束检测

在空中交通管制领域，检测通信结束至关重要。虽然通常可以获取按键通话信号，但有时并非如此。在这种情况下，可以利用 ATCO2 语料库，通过基于声学或基于文本的方法开发通信结束检测系统。基于声学的系统称为语音活动检测系统（VAD），会在空中交通管制通信发送至自动语音识别系统之前执行检测 [76,77]，但可能需要在识别流程中集成一个新的独立模块。基于文本的系统依赖 BERT 等强大的人工智能模型；此前关于空中交通管制的研究 [12] 已表明，这类系统能够有效检测呼号 [78]、指令 [10]，以及自动语音识别系统生成的转录文本中的通信结束信号。

5.3.5. 复诵错误检测

复诵错误检测（RBED）是空中交通管制员的重要职责，因为未被发现的错误可能导致空中交通管理中的危险情况。尽管空中交通管制（ATC）中的通信错误并不常见，但仍可能引发严重的安全问题，有些传输中甚至包含多个错误。先前研究表明，每一百次 [79,80] 至每十六次空中交通管制通信中就可能有一次包含错误 [81]。幸运的是，航路空域的空中交通管制员检测此类错误的能力相当强，研究显示，他们能够检测出多达90%的飞行员复诵错误 [79]。虽然复诵错误总体而言十分罕见，但通过自动复诵错误检测避免哪怕一起事故，也能为保障空中交通管理安全带来重大助益。为支持空中交通管制员完成这项任务，先前的项目采用自动语音识别与理解引擎，从空中交通管制通信中提取高层信息 [32]。在 [82] 中提出了两种复诵错误检测方法。一种系统基于规则，另一种则是数据驱动的序列分类器，采用类似 BERT 的预训练编码器 RoBERTa [57]。在此，输入序列由空中交通管制员和飞行员的话语转录文本拼接而成，中间以特殊分隔符标记 [SEP] 隔开。研究表明，将这两种方法结合起来，在 Isavia 航路空域的真实语音录音中，复诵错误检测率可达81%。研究还介绍了一项概念验证试验，由六名空中交通管制员生成具有挑战性的人工复诵错误样本。

然而，HAAWAII 或 MAL-ORCA 等知名的过往项目有一个主要问题：其数据无法公开共享。相比之下，ATCO2 语料库面向公众开放，例如，*ATCO2-test-set-1h* 测试集可免费访问，从业者也可以参考以往的研究来实现 RBED 模块。

5.3.6. 英语检测

目前，我们已开发并部署了合适的英语检测系统（ELD），用于剔除新采集数据中的非英语话语。我们测试了一种采用 x-vector 提取器、基于声学特征的先进系统。我们还提出了一种自然语言处理方法：利用带有词置信度的自动语音识别输出进行 ELD。最后，我们的实验表明，基于自然语言处理的 ELD 在检测准确率和计算资源方面均优于声学方法。此外，自然语言处理方法可以联合使用多个自动语音识别系统的输出，从而进一步提升结果。在处理流水线中，我们集成了基于自然语言处理的英语检测器，该检测器运行于捷克语和英语自动语音识别系统之上。集成的英语检测器由用于重新加权累积“软”词计数的词频-逆文档频率，以及用于判定英语/非英语的逻辑回归分类器组成 [27]。

我们创建了开发与评估数据集，其中包含来自不同机场的数据、带有各种英语口音的数据，以及英语与当地语言的语码混用数据。这些数据选自我们在表 1 中介绍的 ATCO2语料库。开发集用于估计我们的英语语言检测器（逻辑回归分类器）的模型参数。评估集用于测试。手动标注话语的规则

下文将提及。我们在聆听并标注英语语言检测数据集时，发现了空中交通管制数据的几个有趣特性：

• 噪声条件各不相同。大多数数据都很干净，但有些片段噪声非常大。

• 英语口音浓重。说话人的英语口音差异很大，从母语者（飞行员）到带有国际口音（法语、德语、俄语等）的飞行员和空中交通管制员，再到口音浓重的捷克语飞行员和空中交通管制员。

• 词语和短语混用。例如，捷克语空中交通管制员的词汇表混合了捷克语和英语词汇。他们使用捷克语的标准问候语；如果指令很短，这些问候语可能会占据“英语”句子的相当大一部分。另一方面，他们会在“捷克语”句子中使用许多英语词汇（字母表、某些指令）。此外，他们还使用大量“捷克式英语”词汇。

我们通常根据口述数字所使用的语言，来判断某段空中交通管制员与飞行员之间的通信话语使用何种语言。整个音频录音中的语言必须保持一致。更多详细信息（包括实验结果）请参阅我们先前的研究 [27]。

6. 结论

本文在我们之前的工作 [10] 基础上进行了拓展，并讨论了从ATCO2项目中获得的主要经验教训。ATCO2项目旨在开发一个用于收集、预处理和伪标注空中交通管制通信音频数据的平台。ATCO2拥有超过5000小时的伪标注音频数据，是迄今为止最大的公开空中交通管制数据集。该项目推动了空中交通管制通信的鲁棒自动语音识别和自然语言理解研究。

从 ATCO2 中汲取的主要经验如下：

• **经验教训 1：** ATCO2 的自动转录引擎（参见附录 H）和标注平台（参见附录 I）已证明能够可靠地收集面向空中交通管制通信的大规模音频数据集。

• **教训 2：** 基于以往项目发布的本体，制定了良好的空中交通管制通信标注实践 [32]。我们制作了一份速查表（见附录K），为未来的空中交通管制项目提供指导，并减少转录文本生成过程中的混淆。

• **第3课：** 仅使用 ATCO2 数据集（例如 *ATCO2-PL-500hsetcorpus*）训练自动语音识别系统，也能在 ATCO2 测试集上取得有竞争力的词错误率（见表 4）。该自动语音识别模型在 ATCO2-test-set-1h/ATCO2-test-set-4h 上的词错误率分别可低至 17.9%/24.9%。更重要的是，这些测试集包含带有口音且受噪声干扰的语音，对标准自动语音识别系统来说极具挑战性。

• **第4课：** 空中交通管制监视数据是改善自动语音识别输出的实时信息的理想来源。整合空中监视数据可使呼号词错误率最多绝对降低 11.8%，这意味着在 ATCO2 测试集-4h 上，呼号准确率提升了 62.6%（未进行增强） →82.9%（经 GT 增强），如表5所示。

• **第5课：** ATCO2语料库可用于对空中交通管制通信进行自然语言理解。基于 ATCO2-test-set-4h，已开发出基于BERT的命名实体识别和说话人角色检测模块。这些系统可以从文本输入中检测呼号、指令和数值。此外，还可以基于文本输入检测说话人角色。自然语言理解任务对空中交通管制领域尤其重要，因为这些高层信息可用于协助空中交通管制员完成日常任务，从而降低其总体工作负荷。

总之，从 ATCO2项目中汲取的经验，以及其大规模音频数据库收集和预转录方法，可应用于数据稀缺这一潜在挑战普遍存在的其他领域。

致谢

本文是我们此前工作的延续 [10]，*ATCO2*语料库。ATCO2源于Clean Sky 2联合事业（JU）和EU-H2020的共同贡献，并由Clean Sky 2联合事业（JU）和EU-H2020通过第864702号资助协议提供支持——ATCO2（从空中交通通信中自动收集和处理语音数据）。请访问我们的网站：https://www.atco2.org/。

作者贡献

概念化：Juan Pablo Zuluaga Gomez、Srikanth Madikeri、Igor Szoke、Vincent Lenders、Mickael Rigault 和 Khalid Choukri；数据整理：Juan Pablo Zuluaga Gomez 和 Iuliia Nigmatulina；正式分析：Juan Pablo Zuluaga Gomez 和 Igor Szoke；资金获取：Petr Motlicek；调查：Juan Pablo Zuluaga Gomez；方法论：Juan Pablo Zuluaga Gomez、Iuliia Nigmatulina、Amrutha Prasad 和 Allan Tart；项目管理：Petr Motlicek；资源：Juan Pablo Zuluaga Gomez；软件：Juan Pablo Zuluaga Gomez、Srikanth Madikeri、Allan Tart 和 Igor Szoke；监督：Petr Motlicek、Srikanth Madikeri、Igor Szoke 和 Vincent Lenders；验证：Juan Pablo Zuluaga Gomez、Amrutha Prasad 和 Igor Szoke；可视化：Juan Pablo Zuluaga Gomez 和 Mickael Rigault；撰写——初稿：Juan Pablo Zuluaga Gomez 和 Iuliia Nigmatulina；撰写——审阅与编辑：Juan Pablo Zuluaga Gomez、Iuliia Nigmatulina、Amrutha Prasad、Petr Motlicek、Driss Khalil、Vincent Lenders、Mickael Rigault 和 Khalid Choukri。

7. 参考文献

1. Kleinert, M.; Helmke, H.; Siol, G.; Ehr, H.; Cerna, A.; Kern, C.; Klakow, D.; Motlicek, P.; Oualil, Y.; Singh, M.; 等. 面向不同进近区域的基于助手的语音识别模型半监督适配。载于第37届数字航空电子系统会议（DASC）论文集。IEEE，2018年，第1–10页。2. Kleinert, M.; Helmke, H.; Siol, G.; Ehr, H.; Klakow, D.; Singh, M.; Motlicek, P.; Kern, C.; Cerna, A.; Hlousek, P. 将基于助手的语音识别适配到新领域及其在空中交通管制员中的接受度。载于国际智能人类系统集成会议论文集。Springer，2019年，第820–826页。3. Helmke, H.; Ohneiser, O.; Mühlhausen, T.; Wies, M. 利用自动语音识别降低管制员工作负荷。载于2016年IEEE/AIAA第35届数字航空电子系统会议（DASC）论文集。IEEE，2016年，第1–10页。4. Helmke, H.; Ohneiser, O.; Buxbaum, J.; Kern, C. 利用基于助手的语音识别提高空中交通管理效率。载于第13届美国/欧洲空中交通管理研发研讨会论文集，美国西雅图，2017年。5. Nigmatulina, I.; Zuluaga-Gomez, J.; Prasad, A.; Sarfjoo, S.S.; Motlicek, P. 利用上下文数据的两步法：空中交通通信中的语音识别。载于ICASSP论文集，2022年。6. Kleinert, M.; Helmke, H.; Shetty, S.; Ohneiser, O.; Ehr, H.; Prasad, A.; Motlicek, P.; Harfmann, J. 空中交通管制通信的自动解读：从口语到深入理解其含义的历程。载于2021年IEEE/AIAA第40届数字航空电子系统会议（DASC）论文集。IEEE，2021年，第1–9页。7. Kleinert, M.; Shetty, S.; Helmke, H.; Ohneiser, O.; Wiese, H.; Maier, M.; Schacht, S.; Nigmatulina, I.; Saeed, S.; Sarfjoo, P.M. 通过将自动语音识别与先进的地面活动引导与控制系统集成，为机坪管制员提供支持**2022**。

8. Cordero, J.M.; Dorado, M.; de Pablo, J.M. 空中交通管制环境中的自动语音识别。载于第二届指挥与控制系统自动化应用与理论国际会议论文集，2012年，第46–53页。9. Ferreiros, J.; Pardo, J.; De Córdoba, R.; Macias-Guarasa, J.; Montero, J.; Fernández, F.; Sama, V.; González, G.; 等. 用于空中交通管制终端的语音接口。《航空航天科学与技术》2012年，21，7–15。10. Zuluaga-Gomez, J.; Vesel` y, K.; Szöke, I.; Motlicek, P.; Kocour, M.; Rigault, M.; Choukri, K.; Prasad, A.; Sarfjoo, S.S.; Nigmatulina, I.; 等. ATCO2语料库：用于空中交通管制通信自动语音识别和自然语言理解研究的大规模数据集。arXiv预印本 arXiv:2211.04054，2022年。11. Kocour, M.; Veselý, K.; Szöke, I.; Kesiraju, S.; Zuluaga-Gomez, J.; Blatt, A.; Prasad, A.; Nigmatulina, I.; Motlíˇcek, P.; Klakow, D.; 等. 用于收集和标注空中交通语音通信数据的自动处理流程。《工程会议论文集》2021年，13，8。12. Zuluaga-Gomez, J.; Sarfjoo, S.S.; Prasad, A.; Nigmatulina, I.; Motlicek, P.; Ondre, K.; Ohneiser, O.; Helmke, H. BERTraffic：基于BERT的空中交通管制通信联合说话人角色与说话人变更检测。IEEE口语技术研讨会（SLT），卡塔尔多哈，2023年。13. Kocour, M.; Veselý, K.; Blatt, A.; Gomez, J.Z.; Szöke, I.; Cernocky, J.; Klakow, D.; Motlicek, P. 增强自动语音识别中的上下文信息，以实现空中交通呼号识别。载于Interspeech论文集，2021年，第3301–3305页。https://doi.org/10.21437/Interspeech.2021-1619。14. Nigmatulina, I.; Braun, R.; Zuluaga-Gomez, J.; Motlicek, P. 利用空中监视数据改进空中交通通信中的呼号识别。arXiv预印本 arXiv:2108.12156，2021年。15. Tarakan, R.; Baldwin, K.; Rozen, N. 支持先进空中交通管制员培训的自动模拟飞行员能力。载于第26届国际航空科学大会和第8届AIAA航空航天技术、集成与运营会议论文集，2008年。16. Zuluaga-Gomez, J.; Motlicek, P.; Zhan, Q.; Vesel`y, K.; Braun, R. 空中交通通信自动语音识别基准。载于Interspeech论文集，2020年，第2297–2301页。https://doi.org/10.21437/Interspeech.2020-2173。17. Srinivasamurthy, A.; Motlicek, P.; Himawan, I.; Szaszak, G.; Oualil, Y.; Helmke, H. 通过语义知识提取进行半监督学习，以改进空中交通管制中的语音识别。载于国际语音通信协会第18届年会论文集，2017年。18. Zuluaga-Gomez, J.; Nigmatulina, I.; Prasad, A.; Motlicek, P.; Vesel`y, K.; Kocour, M.; Szöke, I. 上下文半监督学习：利用空中监视数据和未转录空中交通管制数据改进自动语音识别系统的方法。载于Interspeech论文集，2021年，第3296–3300页。https://doi.org/10.21437/Interspeech.2021-1373。19. Zuluaga-Gomez, J.; Vesel`y, K.; Blatt, A.; Motlicek, P.; Klakow, D.; Tart, A.; Szöke, I.; Prasad, A.; Sarfjoo, S.; Kolˇcárek, P.; 等. 自动呼号检测：将空中监视数据与空中交通口语通信相匹配。载于多学科数字出版研究院会议论文集，2020年，第59卷，第14页。20. Godfrey, J. 空中交通管制语料库（ATC0）-LDC94S14A，1994年。21. Šmídl, L.; Švec, J.; Tihelka, D.; Matoušek, J.; Romportl, J.; Ircing, P. 空中交通管制通信（ATCC）语音语料库及其在自动语音识别和文本转语音开发中的应用。《语言资源与评测》2019年，53，449–464。22. Segura, J.; Ehrette, T.; Potamianos, A.; Fohr, D.; Illina, I.; Breton, P.; Clot, V.; Gemello, R.; Matassoni, M.; Maragos, P. HIWIRE数据库：用于驾驶舱通信的嘈杂非母语英语语音语料库。在线。http://www.hiwire. org，2007年。23. Delpech, E.; Laignelet, M.; Pimm, C.; Raynal, C.; Trzos, M.; Arnold, A.; Pronto, D. 真实场景下带法国口音的空中交通管制通信语料库。载于第十一届语言资源与评测国际会议（LREC 2018）论文集，2018年。24. Pellegrini, T.; Farinas, J.; Delpech, E.; Lancelot, F. 空中客车空中交通管制语音识别2018年挑战赛：迈向空中交通管制自动转录和呼号检测。arXiv预印本 arXiv:1810.12614，2018年。25. Graglia, L.; Favennec, B.; Arnoux, A. Vocalise：评估数据链技术对无线电话信道的影响。载于第24届数字航空电子系统会议论文集。IEEE，2005年，第1卷，第5–C页。

26. Lopez, S.; Condamines, A.; Josselin-Leray, A.; O'Donoghue, M.; Salmon, R. 空地通信中英语固定短语和明白语言的语言学分析。《航空运输研究杂志》**2013**，*4*，44–60。27. Szöke, I.; Kesiraju, S.; Novotný, O.; Kocour, M.; Veselý, K.; ˇCernocký, J. 检测空中交通管制语音通信中的英语语音。载于 Interspeech 2021 会议论文集，2021，第 3286–3290 页。https://doi.org/10.21437/Interspeech.2021-1033。28. 国际民用航空组织。《国际民用航空组织用语参考指南》，2020。29. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech：一个基于公有领域有声读物的自动语音识别语料库。载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集。IEEE，2015，第 5206–5210 页。30. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common Voice：一个大规模多语言语音语料库。 *arXiv* 预印本 *arXiv:1912.06670***2019**。31. Godfrey, J.J.; Holliman, E.C.; McDaniel, J. SWITCHBOARD：用于研究与开发的电话语音语料库。载于 IEEE 国际声学、语音与信号处理会议论文集。IEEE计算机学会，1992，第 1 卷，第 517–520 页。32. Helmke, H.; Slotty, M.; Poiger, M.; Herrer, D.F.; Ohneiser, O.; Vink, N.; Cerna, A.; Hartikainen, P.; Josefsson, B.; Langr, D.; 等。用于转录 SESAR 2020 解决方案 PJ. 16-04 中空中交通管制语音指令的本体。载于 IEEE/AIAA 第 37 届数字航空电子系统会议（DASC）论文集。IEEE，2018，第 1–10 页。33. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; 等。Kaldi 语音识别工具包。载于 IEEE 自动语音识别与理解研讨会论文集。IEEE信号处理学会，2011，会议编号 CONF。34. Rigault, M.; Cevenini, C.; Choukri, K.; Kocour, M.; Vesel`y, K.; Szoke, I.; Motlicek, P.; Zuluaga-Gomez, J.P.; Blatt, A.; Klakow, D.; 等。录制空中交通管制语音的法律与伦理挑战。载于第 13 届语言资源与评测会议中的人类语言技术伦理与法律问题及语言资源中敏感数据多语言去标识化研讨会论文集，2022，第 79–83 页。35. Jurafsky, D.; Martin, J.H. 语音与语言处理（第2版）；Prentice-Hall, Inc.：美国新泽西州萨德尔里弗，2009。36. Wang, D.; Wang, X.; Lv, S. 端到端自动语音识别概述。《*Symmetry*》**2019**，*11*，1018。37. Kingsbury, B. 基于格的神经网络声学建模序列分类准则优化。载于 2009 年 IEEE 国际声学、语音与信号处理会议论文集。IEEE，2009，第 3761–3764 页。38. Povey, D.; Peddinti, V.; Galvez, D.; Ghahremani, P.; Manohar, V.; Na, X.; Wang, Y.; Khudanpur, S. 基于无格最大互信息进行纯序列训练的自动语音识别神经网络。载于 Interspeech，2016，第 2751–2755 页。39. Graves, A.; Jaitly, N. 借助循环神经网络实现端到端语音识别。载于第 31 届国际机器学习会议（ICML 2014）论文集，中国北京，2014 年 6 月 21–26 日。JMLR.org，2014，第 32 卷，*JMLR*研讨会与会议论文集，第 1764–1772 页。40. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. 基于注意力的语音识别模型。载于神经信息处理系统进展会议论文集，2015 年 12 月 7–12 日，加拿大魁北克省蒙特利尔，2015，第 577–585 页。41. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0：一种用于语音表征自监督学习的框架。载于神经信息处理系统进展会议论文集，2020 年 12 月 6–12 日，线上，2020。42. Chen, S.; Wang, C.; Chen, Z.; 等。Wavlm：用于全栈语音处理的大规模自监督预训练。 *arXiv* 预印本**2021**，*abs/2110.13900*。43. Oord, A.v.d.; Li, Y.; Vinyals, O. 通过对比预测编码进行表征学习。 *arXiv* 预印本 **2018**

, *abs/1807.03748* .

44. Baevski, A.; Schneider, S.; Auli, M. vq-wav2vec：离散语音表示的自监督学习。载于第八届国际学习表征会议论文集，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。OpenReview.net，2020年。45. Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; von Platen, P.; Saraf, Y.; Pino, J.; 等。Xls-r：大规模自监督跨语言语音表征学习。*arXiv*预印本**2021**，*abs/2111.09296*。46. Zuluaga-Gomez, J.; Prasad, A.; Nigmatulina, I.; Sarfjoo, S.; Motlicek, P.; Kleinert, M.; Helmke, H.; Ohneiser, O.; Zhan, Q. 预训练的 Wav2Vec2.0 在领域偏移自动语音识别中的表现如何？一项针对空中交通管制通信的全面基准测试。*IEEE*口语技术研讨会（*SLT*），卡塔尔多哈**2023**。47. Hall, K.; Cho, E.; Allauzen, C.; Beaufays, F.; Coccaro, N.; Nakajima, K.; Riley, M.; Roark, B.; Rybach, D.; Zhang, L. 基于组合的即时重评分，用于显著的 n 元语法偏置**2015**。48. Aleksic, P.; Ghodsi, M.; Michaely, A.; Allauzen, C.; Hall, K.; Roark, B.; Rybach, D.; Moreno, P. 将上下文信息引入 Google 语音识别**2015**。49. Serrino, J.; Velikovich, L.; Aleksic, P.S.; Allauzen, C. 自动语音识别中格外命名实体的上下文恢复。载于 Interspeech 论文集，2019年，第3830–3834页。50. Chen, Z.; Jain, M.; Wang, Y.; Seltzer, M.L.; Fuegen, C. 使用类别语言模型和令牌传递解码器的端到端上下文语音识别。载于 ICASSP 2019——2019年 IEEE国际声学、语音与信号处理会议（ICASSP）论文集。IEEE，2019年，第6186–6190页。51. Yadav, V.; Bethard, S. 深度学习模型中命名实体识别最新进展综述。载于第二十七届国际计算语言学会议论文集，2018年，第2145–2158页。52. Sharma, A.; Chakraborty, S.; Kumar, S.; 等。自然语言处理中命名实体识别：系统综述。载于第二届计算智能博士生研讨会论文集。Springer，2022年，第817–828页。53. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. 几乎从零开始的自然语言处理。机器学习研究杂志**2011**，*12*，2493–2537。54. Birjali, M.; Kasri, M.; Beni-Hssane, A. 情感分析综合综述：方法、挑战与趋势。*Knowledge-Based Systems***2021**，*226*，107134。55. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. 注意力机制就是一切。神经信息处理系统进展**2017**，*30*。56. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT：用于语言理解的深度双向 Transformer 预训练。*arXiv*预印本 *arXiv:1810.04805* **2018**。57. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa：一种稳健优化的 BERT 预训练方法。*arXiv*预印本 *arXiv:1907.11692***2019**。58. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; 等。Transformers：自然语言处理领域的最新技术。载于2020年自然语言处理实证方法会议论文集：系统演示，2020年，第38–45页。59. Lhoest, Q.; del Moral, A.V.; Jernite, Y.; Thakur, A.; von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; 等。Datasets：一个用于自然语言处理的社区库。载于2021年自然语言处理实证方法会议论文集：系统演示，2021年，第175–184页。60. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout：一种防止神经网络过拟合的简单方法。机器学习研究杂志**2014**，*15*，1929–1958。61. Hendrycks, D.; Gimpel, K. 高斯误差线性单元（GELU）。*arXiv*预印本 *arXiv:1606.08415* **2016**。62. Pascanu, R.; Mikolov, T.; Bengio, Y. 关于循环神经网络训练的困难性。载于国际机器学习会议论文集。PMLR，2013年，第1310–1318页。63. Loshchilov, I.; Hutter, F. 解耦权重衰减正则化。载于国际学习表征会议论文集，2019年。64. He, Z.; Wang, Z.; Wei, W.; Feng, S.; Mao, X.; Jiang, S. 深度学习模型中序列标注最新进展综述。*arXiv*预印本 *arXiv:2011.06727***2020**。

65. Zhou, C.; Cule, B.; Goethals, B. 基于模式的序列分类。*IEEE*知识与数据工程汇刊**2015**，*28*，1285–1298。66. He, P.; Gao, J.; Chen, W. DeBERTaV3：通过采用ELECTRA风格的预训练和梯度解耦嵌入共享改进DeBERTa。*arXiv*预印本 *arXiv:2111.09543***2021**。67. Prasad, A.; Zuluaga-Gomez, J.; Motlicek, P.; 等。用于空中交通管制语音识别的基于语法的说话人角色识别。SESAR联合事业组织，2022。68. Zuluaga-Gomez, J.; Prasad, A.; Nigmatulina, I.; Motlicek, P.; Kleinert, M. 用于培训空中交通管制员的虚拟模拟飞行员智能体。*arXiv*预印本 *arXiv:2304.07842***2023**。69. Madikeri, S.; Bourlard, H. 用于说话人分离的基于滤波器组斜率的特征。载于 ICASSP 论文集。IEEE，2014，页码 111–115。70. Sell, G.; Snyder, D.; McCree, A.; Garcia-Romero, D.; Villalba, J.; Maciejewski, M.; Manohar, V.; Dehak, N.; Povey, D.; Watanabe, S.; 等。说话人分离并非易事：JHU团队在首届DIHARD挑战赛中的经验与教训。载于 Interspeech 论文集，2018，页码 2808–2812。71. Valente, F.; Motlicek, P.; Vijayasenan, D. 会议录音的变分贝叶斯说话人分离。载于2010年IEEE国际声学、语音与信号处理会议论文集。IEEE，2010，页码 4954–4957。72. Landini, F.; Profant, J.; Diez, M.; Burget, L. 说话人分离中的x向量序列贝叶斯隐马尔可夫模型聚类（VBx）：理论、实现及标准任务分析。计算机语音与语言**2022**，*71*，101254。73. Fujita, Y.; Kanda, N.; Horiguchi, S.; Nagamatsu, K.; Watanabe, S. 采用无排列目标函数的端到端神经说话人分离。*arXiv*预印本 *arXiv:1909.05952***2019**。74. Fujita, Y.; Kanda, N.; Horiguchi, S.; Xue, Y.; Nagamatsu, K.; Watanabe, S. 采用自注意力的端到端神经说话人分离。载于2019年IEEE自动语音识别与理解研讨会（ASRU）论文集。IEEE，2019，页码 296–303。75. Ryant, N.; 等。第二届DIHARD说话人分离挑战赛：数据集、任务与基线系统。载于2019年Interspeech会议论文集，2019，页码 978–982。76. Sarfjoo, S.S.; Madikeri, S.; Motlicek, P. 基于多语种语音识别系统的语音活动检测。*arXiv*预印本 *arXiv:2010.12277***2020**。77. Ariav, I.; Cohen, I. 一种使用WaveNet编码器和残差网络的端到端多模态语音活动检测方法。*IEEE*信号处理精选主题期刊**2019**，*13*，265–274。78. Lin, Y. 空中交通管制中的口语指令理解：挑战、技术与应用。航空航天**2021**，*8*，65。79. Cardosi, K.M. 航路管制员与飞行员语音通信分析。技术报告，约翰·A·沃尔普国家交通系统中心，马萨诸塞州剑桥，1993。80. Cardosi, K.M.; 等。塔台（本场）管制员与飞行员语音通信分析。技术报告，美国联邦航空管理局，1994。81. Prinzo, O.V. 空中交通管制信息复杂度和信息长度对飞行员复诵表现的计算及影响。《行为测量》*2008***2008**，*188*。82. Helmke, H.; Ondˇrej, K.; Shetty, S.; Arilíusson, H.; Simiganoschi, T.; Kleinert, M.; Ohneiser, O.; Ehr, H.; Zuluaga-Gomez, J.; Smrz, P. 通过自动语音识别与理解进行复诵错误检测——HAAWAII项目针对Isavia航路空域的成果。载于第12届SESAR创新日论文集。SESAR联合事业组织，2022。

附录H 自动转录引擎

本附录详细介绍了我们如何收集音频和元数据，从而构建*ATCO2*语料库。我们主要依赖自动转录引擎，详情见第3.2节。自动转录引擎以可扩展的云服务形式实现。它通过API与其他服务（或合作伙伴）进行通信。该服务旨在处理由数据提供者产生的大量数据流。[15]

数据通过调用 API请求并提供作业设置 JSON文件，由 OSN[16] 服务器推送到此服务。请求被接受后，设置参数将被处理，作业会存储在内部队列中等待处理。用户（在此情况下为 OSN）可能能够调整设置，并影响处理流水线及结果。具体而言：

- 音频输入格式选项；
- 音频过长时的拒绝阈值；
- 过短音频的拒绝阈值；
- 过于嘈杂音频的拒绝阈值；
- 非英语音频的拒绝阈值；
- 切换自动语音识别器的语言。

其中大多数实际上因安全原因而被禁用（以免中断处理流水线），但如有需要，可随时轻松启用。整体数据流

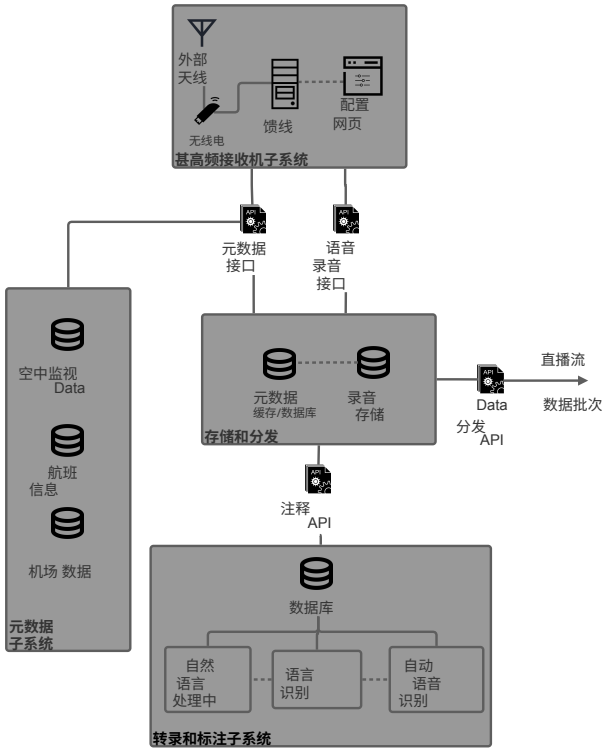


图 A13. *ATCO2*通信架构总体示意。

模型如图所示 A13。通过 SpokenData 的 API 接受的任何新作业（请求对录音进行完整的自动标注）[17] 都会由主处理节点处理。作业会加入工作负载管理器队列。只要有空闲的处理槽位，作业就会提交给处理服务器（即工作节点）。随后，主处理节点会通过调用回调，向 OSN 服务器告知作业状态。

[15] 作为空中交通管制语音和相关空中交通管制数据（监视数据）“馈线”的爱好者。请参见第 4.2节。[16] OpenSky Network：https://opensky-network.org/。[17] 工业合作伙伴：https://www.spokendata.com/atco2。

附录 I 标注平台：数据流

数据（用于人工标注的录音）生命周期分为 4 个主要状态：

录音从转录引擎推送到标注平台时，其初始状态会设为“已排队”，且尚未处理。录音会加入标注任务队列，并立即对所有标注员可见。该队列显示在开放任务界面中。标注员可以与队列互动——收听录音，并选择要标注的录音。在以下情况下，处于此状态的录音可能会从队列中移除：录音时间较久，无人有意愿进行标注；或有 3 名标注员对录音点了踩。移出的录音将在 7 天后删除。

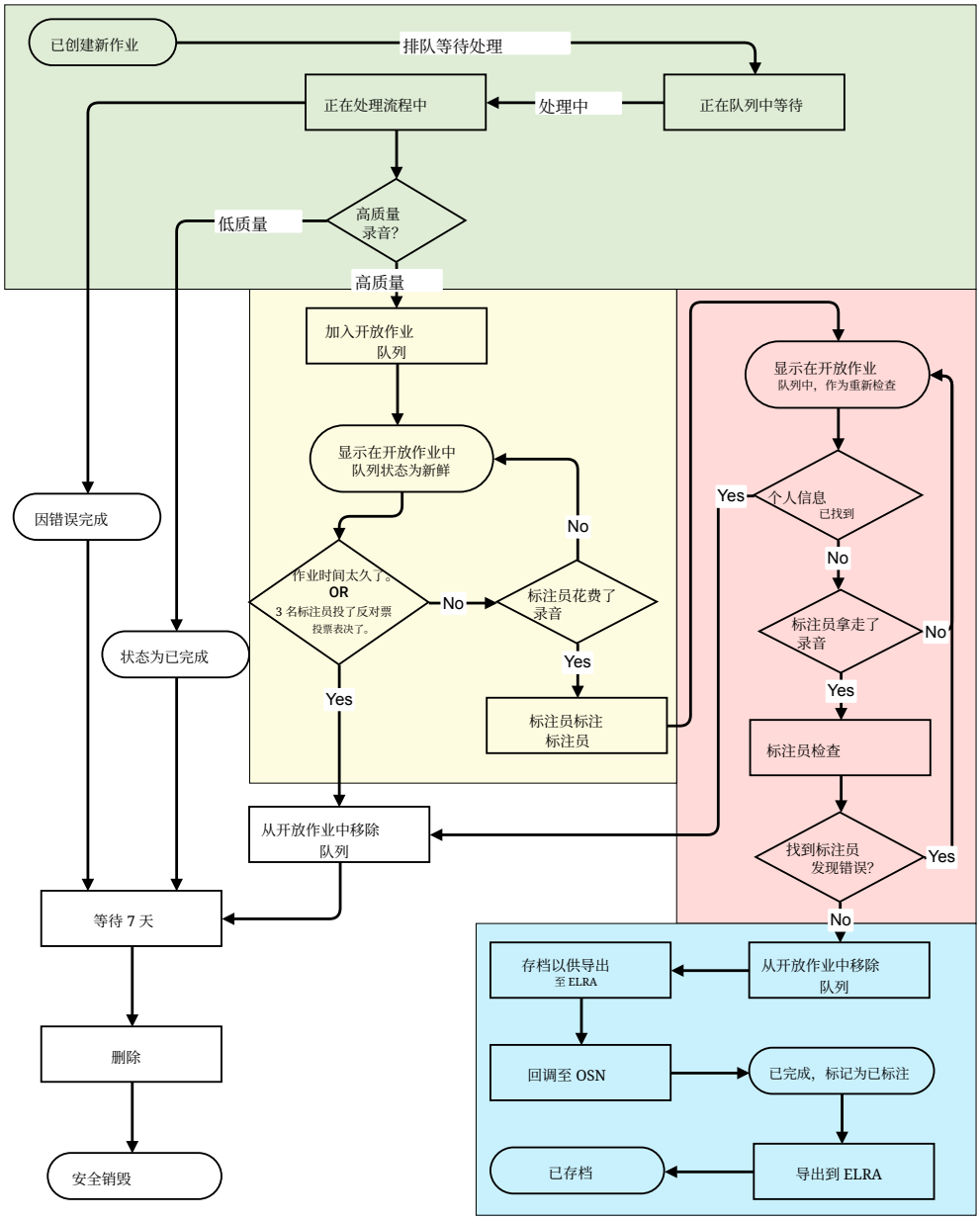
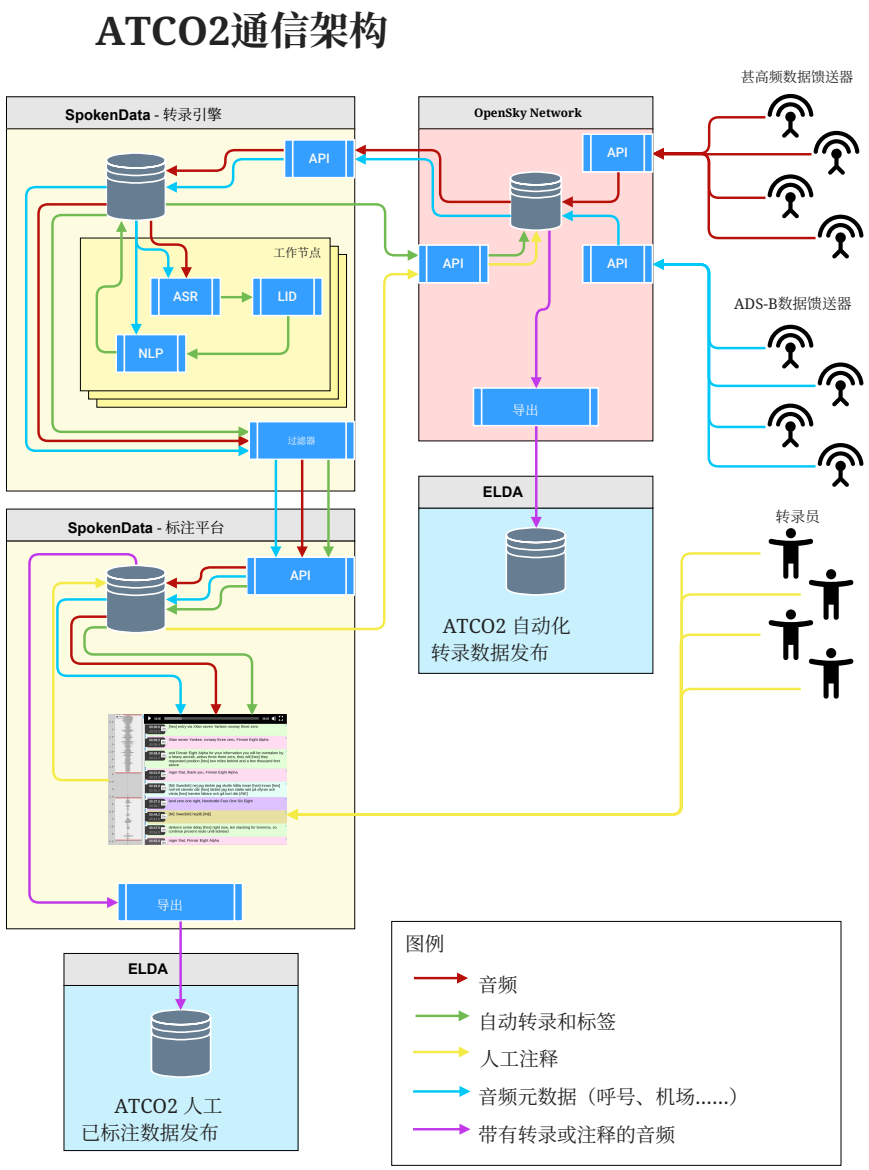


图 A14。 标注平台的数据流。标注平台中数据流（生命周期）的示意图。转录引擎以绿色表示。已排队且未处理的状态以黄色表示。已排队且已标注的状态以红色表示。已标注状态以蓝色表示。其余部分（白色）表示安全销毁状态。

用户选择要标注的录音、修改自动标注并保存后，该录音将被设为“已排队”和“已标注”。此状态可防止录音从队列中移除或被删除。此外，开放任务界面会将其标记为“待重新检查”，以告知其他标注员该录音已被修改（已标注），他们应重新检查标注是否正确，而不是从头开始标注。如果任何标注员指出录音中存在个人信息（通过“匿名化”标签），该录音将从队列中移除并被删除。

下一个状态为已标注。如果录音成功通过重新检查，则该录音将被视为已标注，且标注结果为最终版本。该录音将从开放任务队列中移除，并放入已完成录音标注的堆栈中。该堆栈会定期导出至 ELRA，以便进一步打包并分发给相关社区。此状态还会触发向 OSN平台发送回调，通知其人工标注已完成，并且可以下载转录文本。录音导出至 ELRA 后，我们将其状态设置为 已完成。在此状态下，可以对录音进行存档或删除。详细的数据流示意图如图 A14 所示。

附录 J 通信模式ATCO2 项目期间开发的通信模式如图所示 A15。



图A15. *ATCO2*整体通信架构。

附录 K 标注速查表图 A16 展示了 ATCO2项目制定的标注速查表。

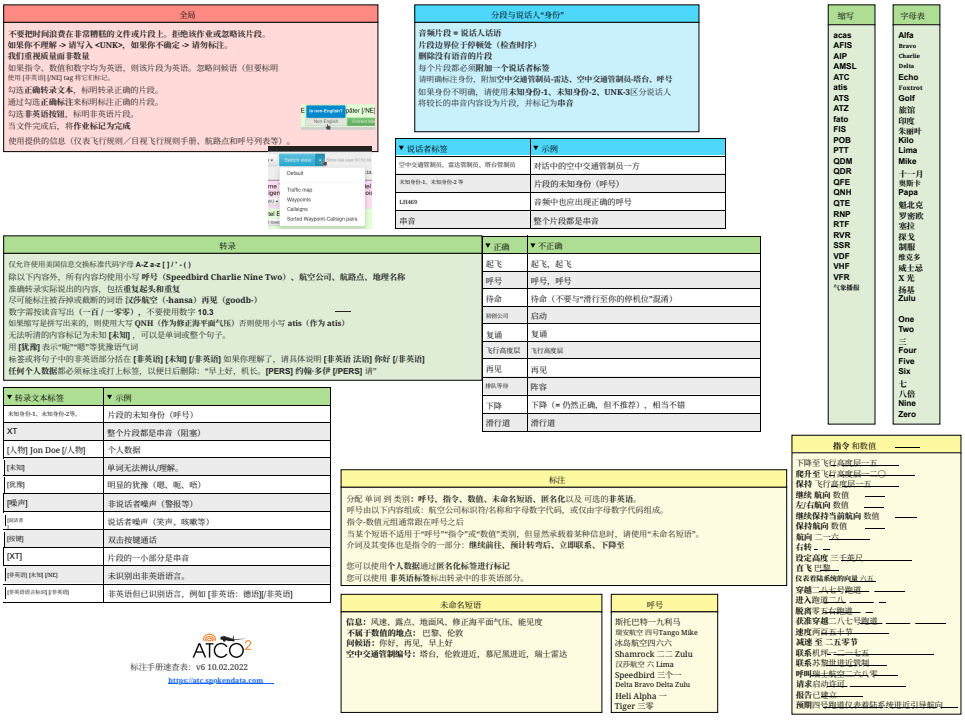


图 A16。 速查表。 在 ATCO2语料库标注过程中编制的文档。此文档可用于标注来自不同机场的空中交通管制通信数据。