arXiv:2603.01894v1 [cs.SD] 2 Mar 2026

VietSuperSpeech：面向聊天机器人、客户支持和呼叫中心应用的用于自动语音识别微调的大规模越南语对话式语音数据集

LoanDo* ThanhNgocNguyen ThanhPham FPTUniversity ICMS RMITUniversityVietnam thnguyen@icms.edu.au thanh.pham@rmit.edu.vn

VinhDo HienNguyen NGHIStudio PhuongHai JSC ddvinh1@gmail.com hien.nm@phuonghai.com

Charlotte Nguyen NGHI Studio charlottenguyen1501@gmail.com

2026年2月

摘要

我们介绍 **VietSuperSpeech**，这是一个大规模越南语自动语音识别（自动语音识别）数据集，包含 52,023 个音频文本对，总时长达 267.39 小时，尤为侧重随意的对话语音。与现有越南语自动语音识别语料库主要包含朗读语音、新闻旁白或有声读物内容不同，VietSuperSpeech 的素材来自四个可公开访问的 YouTube频道，涵盖日常对话、个人视频博客、海外越南社群对话和非正式评论——这些正是现实中的聊天机器人、客户支持、呼叫中心和热线部署所涉及的语音风格。所有音频均统一为 16 kHz 单声道 PCM WAV，并切分为 3–30 秒的话语。转写文本通过伪标签生成，使用 Zipformer-30M-RNNT-6000h 模型 [Nguyen，该模型 2025]通过 Sherpa-ONNX 部署，并在 6,000 小时越南语语音上进行预训练。经过质量筛选后，数据集以固定随机种子划分为 46,822 个训练样本（240.67 小时）和 5,201 个开发集/测试集样本（26.72 小时）。每条话语的文本平均为 266 个字符，总计 1,380 万个带有完整变音符号的越南语字符。我们证明，VietSuperSpeech 填补了越南语自动语音识别生态系统中的关键空白：尽管 VLSP2020 [VLSP组委会、 2020], VIET_BUD500 [Pham 等人，2024a],VietSpeech [Pham 等人，2024b], FLEURS [Conneau等人、 2022], VietMed [Le-Duc、 2024],Sub-GigaSpeech2-Vi [Yang 等人、 2024], viVoice [Le 等人，2024b], 和 Sub-PhoAudioBook [Vu 等人 2025] 等语料库广泛覆盖了正式语体和朗读语音，但没有一个专门针对对话式人工智能应用不可或缺的随意、自发语体。VietSuperSpeech 已公开发布于 https://huggingface.co/datasets/thanhnew2001/VietSuperSpeech。

*通讯作者：dttloan.ute@gmail.com

1 引言

与英语等资源丰富的语言相比，越南语自动语音识别面临若干独特挑战。越南语是一种声调语言，具有六个词汇声调 [Luong and Vu、复杂的音节结构，以及北部、中部和南部方言之间显著的音系变异。随着越南语语言技术日趋成熟，自动语音识别的实际应用重点正迅速从广播转写转向对话应用：语音聊天机器人、客户支持自动化、联络中心分析，以及热线交互式语音应答系统。这些应用场景对训练数据提出了本质上不同的一系列要求，而现有的越南语自动语音识别语料库现状尚无法充分满足这些要求。

对话语音识别的差距。现有越南语自动语音识别语料库主要涵盖正- 式 或朗读 语体：VIVOS [Luong and Vu, 2016]完全由朗读语音构成；FOSD [FPT技术研究院， 2020] 由录音棚录制；Sub-PhoAudioBook [Vu 等人， 2025]是有声读物朗读材料；FLEURS [Conneau等人，2022]则由平行译文朗读而成。即便是范围更广的语料库——VIET_BUD500 [Pham 等人，2024a], Sub-GigaSpeech2-Vi [Yang 等人，2024], 以及 VietSpeech [Pham 等人，2024b]——其数据也汇集自不同来源，却未专门致力于捕捉真实客户互动中典型的自然、非正式语体。因此，仅使用这些语料库训练或评估的自动语音识别模型，可能难以泛化至呼叫中心和聊天机器人音频中常见的不流畅现象、口语表达、方言混用、语音弱化及轮次交替模式。

VietSuperSpeech。为弥补这一空白，我们推出了**VietSuperSpeech**，这是一个包含267小时越南语语音的语料库，素材来自以轻松交谈为主的YouTube频道，内容涵盖个人视频博客、非正式讨论、海外越南人社区交流以及日常评论。这些频道自然呈现了已部署的对话式人工智能系统中常见的非正式语体、自发性言语不流畅、声学变异和说话人多样性。结合采用最先进的Zipformer-30M-RNNT-6000h模型的可扩展伪标签生成流程 [Nguyen，2025], VietSuperSpeech提供了一项可复现的开放获取资源，专门适用于为对话式越南语应用微调自动语音识别模型。

本研究的主要贡献如下：

1. 一个包含 52,023 条话语、总时长为 267.39 小时的越南语语音数据集，专门为**随意、对话式语音**精心构建，涵盖个人视频博客、非正式评论和海外侨民社区对话；2. 明确面向**聊天机器人、客户支持、呼叫中心和热线**自动语音识别微调的应用场景，并通过系统分析对话语音与正式语音在声学和语言学特性上的差异提供支持；3. 使用 Sherpa-ONNX 和 Zipformer-30M-RNNT-6000h 模型构建可扩展、可复现的伪标签生成流程；4. 提供经过完整预处理的语料库（16 kHz 单声道 WAV，片段时长为 3–30 秒），并按随机种子固定 89/11 的训练集/开发集划分；5. 在 HuggingFace 上公开发布，以支持可复现研究和工业级微调。

2 背景与动机

2.1 部署系统中的对话式自动语音识别

现代语音人工智能应用——从面向客户的聊天机器人，到呼叫中心转录系统和智能交互式语音应答热线——所处理的语音与大多数学术语音识别基准所依据的语音材料有显著差异 [Shriberg, 1994年, Ward, 2019]。对自动语音识别系统构成挑战的对话语音主要特征包括：

- **不流畅现象**：填充停顿（例如越南语犹豫标记、中途改口、重复和自我修正） [Shriberg, 1994];
- **语音弱化**：非正式越南语口语中常见的快速语流现象、元音弱化、辅音脱落和音节省略；
- **口语词汇**：正式语料库中缺乏的俚语、缩略语、借词、与英语的语码转换以及语域特定词汇；
- **声学变异**：非录音棚麦克风、背景噪声（咖啡馆、街道、住宅）、多位同时说话者以及说话距离各异；
- **地区及侨民口音**：对于海外越南社群而言尤为重要，这些社群构成了越南语商业服务客户群体中的重要部分。

在这些场景中部署基于正式语料或朗读语音语料库训练的自动语音识别模型，却不进行领域自适应时，相较于清洁语音基准，词错误率（WER）往往会大幅上升 [Yang 等，2024, Radford 等人， 2023]。VietSuperSpeech 旨在作为领域适配语料库，弥合这一差距。

2.2 越南语自动语音识别数据集

越南语语音识别的发展得益于日益丰富的公开语料库，下面我们将对其进行综述，并特别关注其对话式自动语音识别的适用性。

VIVOS。作为最早开放获取的越南语自动语音识别语料库之一，VIVOS [Luong and Vu, 2016]提供了 15 小时的朗读语音，录制自 65 位身处安静录音室环境中的说话者。它一直是越南语语音识别评测的标准基准，但由于语料内容为朗读语音，不适用于会话领域适应。

VLSP 挑战赛语料库（2020年–2023年）。VLSP 联盟组织了年度语音识别共享任务。VLSP2020 挑战赛 [VLSP组委会， 2020] （VinBigData-VLSP2020）提供了约 250 小时的转录语音，其中 100 小时可公开获取，包括约 20 小时的朗读语音，以及从开放来源抓取并经人工转录、准确率达 96% 的 80 小时自发语音。VLSP2021 挑战赛 [VLSP组委会，2021] 提供了约 280 小时带标签的通用领域数据，以及供半监督赛道使用的 400 小时未标注音频。VLSP2023 挑战赛 [VLSP组委会， 2023]进一步拓展了这些工作，其中包括基于投票的伪标签赛道（VLSP2023-voting-伪标注），该赛道通过多个语音识别系统的输出达成共识来生成标签，为伪标签质量控制提供了重要先例。尽管规模庞大，VLSP 语料库仅向注册参与者开放，且内容偏向正式及脚本化内容。

FPT开放语音数据集（FOSD / FPTVIET）。FPT开放语音数据集 [FPT技术研究院， 2020] 是一个公开可用的朗读语音语料库，录制于干净的录音室环境中，可作为干净语音基线，但无法代表对话式部署场景。

FLEURS。 FLEURS [Conneau等人 2022] 构建了一个涵盖102种语言的多语言基准，该基准建立在FLoRes-101翻译基准之上，其中包含约12小时的越南语朗读语音。它提供了高质量的多语言评估基线，但未涵盖对话语音。

VietMed。 Le-Duc [2024] 介绍了VietMed，这是一个越南语医疗领域自动语音识别语料库，包含16小时带标注的医疗语音、1,000小时未标注的医疗语音，以及1,200小时未标注的通用语音，于LREC-COLING 2024发表。VietMed展示了领域专用语料库在专业应用（临床场景）中的价值，但其会话语体局限于医患互动，而非开放领域随意交谈。

VIET_BUD500（BUD500）。 BUD500 [Pham等人，2024a] 是一个500小时的多样化越南语语料库，涵盖播客、旅行、书籍、食品及其他体裁，覆盖越南北部、南部和中部口音。尽管其内容比朗读语音语料库更为广泛，但播客内容往往偏向半正式独白，而非自发互动对话。

VietSpeech。 Pham 等人 [2024b] 发布了 VietSpeech，这是一个来自多种社交媒体和在线来源、包含 1,100 小时语音的越南语社交语音数据集。其广泛的领域覆盖范围和多样的口音使其成为一项宝贵的大规模资源，但由于缺乏对话式侧重，非正式双向对话并未得到系统性呈现。

Sub-GigaSpeech2-Vi。 GigaSpeech 2 [Yang 等人， 2024]是一个从 YouTube 汇集而成、面向低资源语言的大规模多语言自动语音识别语料库。越南语子集包含约 7,324 小时原始音频，经基于 Noisy Student Training 的筛选后精炼至约 6,039 小时。作为规模最大的伪标注越南语语料库，它提供了广泛的覆盖范围，但并非专为会话语体整理。

viVoice。 Le 等人 [2024b] 发布了 viVoice，这是首个公开可用的大规模越南语文本转语音数据集，包含来自 186 个 YouTube 频道的 1,017 小时以上语音。转写文本由 Whisper Large-V3 生成。虽然它也被用于自动语音识别，但其主要整理标准是文本转语音适用性，因此优先选择清晰、富有表现力的叙述，而非自发随意交谈。

Sub-PhoAudioBook. PhoAudiobook [Vu 等人， 2025] 是一个包含 941 小时高质量越南语有声读物的语料库，录制者来自 735 位讲述者，由 MovianAI 于 ACL2025 发布。其清晰、吐字分明的讲述方式能最大限度地提升文本转语音和零样本合成的性能，但在声学特征上与呼叫中心和聊天机器人的会话语体相距最远。

CommonVoice. Mozilla CommonVoice 的越南语部分 [Ardila 等人， 2020]提供了约 17 小时的众包验证语音。尽管涵盖了多样化说话者，但朗读句子采集方法无法捕捉自然对话动态。

PhoWhisper. PhoWhisper 并非数据集，但它 [Le 等人，2024a] 使用 844 小时的越南语语音对 Whisper [Radford 等人， 2023] 进行了微调，并在 VIVOS、VLSP2020-T1 和 VLSP2020-T2 上建立了当前最先进的基准。PhoWhisper 的结果凸显了领域内数据的重要性：在正式基准上的表现并不能保证在对话场景部署中具备稳健性。

摘要。表 2 展示了 VietSuperSpeech 在现有领域中的定位。据我们所知，现有开放获取的越南语自动语音识别语料库中，没有任何一个同时满足以下条件：(a) 规模足够大（>100小时）；(b) 专为日常对话语音精心构建。VietSuperSpeech 首次兼顾了这两点。

2.3 用于自动语音识别数据创建的伪标签

伪标签是指使用自动语音识别模型自动转录未标注音频，这是一种无需耗费高昂成本进行人工标注即可扩充语料库的实用策略 [Zavaliagkos 和 Colthurst，1998，Radford 等，2023，Gandhi 等人，2023，Yang 等人， 2024]。一种稳健的变体是基于投票的伪标签 [VLSP组委会， 2023], ，即通过共识协调多个自动语音识别系统生成的转录文本。伪标签的质量在很大程度上取决于教师模型在目标领域中的准确率。

2.4 神经转导器与 Zipformer

神经转导器（RNN-T） [Graves， 2012] 已成为流式处理和离线推理中占主导地位的自动语音识别范式。Conformer [Gulati 等人， 2020] 将卷积与自注意力相结合，实现了强大的序列建模能力。Zipformer [Yao 等人， 2024] 进一步通过类 U-Net 多速率架构和 ScaledAdam 优化器提升了编码器效率，以更少的参数实现了更优的性能。Nguyen [2025] 将 Zipformer 架构应用于越南语，并构建了一个综合的自动语音识别与情感识别流程，使用从多个公开语料库（VLSP2020、VLSP2021、VLSP2023、VIET_BUD500、VietSpeech、FLEURS、VietMed、Sub-GigaSpeech2-Vi、viVoice 和 Sub-PhoAudioBook）汇集的约 6,000 小时越南语语音训练 Zipformer-30M-RNNT-6000h 模型。该模型在 VLSP2020 和 VLSP2023 基准测试中取得了当前最佳的词错误率：在 VLSP2020-Test-T1 上为 12.29%，在 VLSP2023-PublicTest 上为 10.40%，超越了参数量多出 3–50× 倍的模型。由于其准确性、效率（在中央处理器上处理 12 秒音频仅需 0.3 秒推理时间）以及广泛的越南语训练覆盖范围，该模型被选为 VietSuperSpeech 的伪标签骨干模型。

3 数据集构建

3.1 数据收集：聚焦对话式内容

VietSuperSpeech 的音频数据来自四个可公开访问的 YouTube 频道，这些频道经过专门筛选，以最大限度地涵盖随意、非正式的对话式越南语语音。这使我们的数据整理理念有别于以往的越南语自动语音识别语料库，后者通常取材于新闻广播、有声书或播客。这四种频道类型分别是：

1. **个人视频博客频道**：创作者以直接、随意的表达方式记录日常生活、分享感想与观点。语音自然随意，包含不流畅现象和口语表达，即使采用独白形式，也体现出自然的轮流发言线索。这种语体与用户在聊天机器人交互中的说话方式非常接近。

2. **海外越南裔社群频道**：由海外越南侨民社群制作的内容，包括非正式对话、社群讨论和文化评论。这些说话者带有受北部口音影响、同时融合了地区特色的口音，偶尔会在越南语与英语或法语之间进行语码转换；他们与海外服务系统互动时所呈现的对话模式，与国际客户支持服务高度相关。

3. **非正式评论与反应频道**：创作者以对话式风格对新闻、娱乐或社群活动作出反应。自发且未经脚本编排的特点，带来了自然的韵律模式、犹豫和口语词汇，真实反映了现实生活中的口语互动。

4. **新闻与讨论混合频道**：少量内容来自将正式新闻播报与非正式小组讨论或问答环节相结合的频道，提供少量半正式语音，以避免语体过度专门化。

这种刻意筛选形成了一个语料库，其主导语体是轻松、对话式的越南语，有别于大多数同类资源所具有的正式或朗读语体。其内容反映了越南语客户支持专员、聊天机器人或呼叫中心交互式语音应答系统在实际运行中会遇到的语音。

3.2 音频预处理

所有收集到的音频均经过标准化预处理流程：

1. **格式规范化　　处理**：音频转换为 16 kHz、单声道、16 位深度的 PCM WAV 格式　　；

2. **语音活动检测（VAD）**：识别静音区域和非语音时段，并据此确定分段边界；

3. **分段**：将音频流切分为时长为 3–30 秒的话语。该时间窗口既适用于简短的对话轮次（问题、肯定回答、简短回应），也适用于较长的非正式叙述片段，同时排除缺乏足够声学上下文的极短片段，以及混合了不同声学条件的过长片段。

3.3 伪标签生成流程

使用 Zipformer-30M-RNNT-6000h 模型 [Nguyen,2025]通过 Sherpa-ONNX 自动生成转写文本。[1] 该模型在约 6,000 小时的越南语语音上进行了预训练，并可通过 https://huggingface.co/hynt/Zipformer-30M-RNNT-6000h 公开获取。对所有分段后的音频进行批量推理，并对生成的假设结果执行以下质量控制步骤：

- **长度过滤**：删除转录文本少于 10 个字符或超过 1,000 个字符的话语；

- **字符级过滤**：含有较高比例的非越南语字符或特殊符号的片段会被丢弃；

[1] https://github.com/k2-fsa/sherpa-onnx

• **基于置信度的过滤**：根据模型内部解码器得分，排除低置信度假设；

• **人工抽检**：从不同信道类型的转写文本中分层随机抽取样本进行人工审查，以验证整体转写质量，并确认准确保留了随意对话语体。

所有转写文本均保留完整的越南语变音标记，以保存对下游语言建模和词汇消歧至关重要的声调信息。

3.4 数据集划分

质量控制完成后，使用固定随机种子对数据集进行划分：

表 1：VietSuperSpeech 数据集统计信息。

划分	话语	时长（小时）	比例
训练集	46,822	240.67	89.0%
开发集/测试集	5,201	26.72	11.0%
总计	**52,023**	**267.39**	**100%**

4 数据集分析

4.1 与现有语料库的比较

表 2 全面比较了 VietSuperSpeech 与现有公开越南语自动语音识别数据集。对话式 一栏明确标示了各语料库针对自然、自发的随意口语的程度——这是 VietSuperSpeech 的关键创新维度。

4.2 随意语音的声学特性

VietSuperSpeech 的会话语体呈现出多种可测量的声学特性，使其有别于正式语音语料库：

语速变化。 对话语音的语速变化幅度比朗读语音更大，这反映了自然的轮次交替模式、强调表达和即兴规划。VietSuperSpeech 中的个人视频博客和海外侨民社区内容，涵盖了真实客户互动中出现的各种语速。

不流利现象密度。 越南语会话语音中常有填充词、中途改口和自我修正。这些词项在朗读语音和有声读物语料库中明显不足 [Shriberg, 1994]但在 VietSuperSpeech 中却随处可见，因此该语料库对于训练抗不流利现象的自动语音识别系统具有重要价值。

声学环境多样性。 内容涵盖家庭录音环境、户外场景、咖啡馆和办公室背景音。这种多样性近似模拟了客户服务通话中的声学条件，此类通话的音频质量不受控制。

表 2：公开越南语自动语音识别数据集对比。“伪” = 伪标注；“人工”= 人工核验；“众包” = 众包。“对话式”列表示该语料库是否专门面向适用于聊天机器人和呼叫中心自动语音识别的随意/自发语音语体。

数据集	小时	转录	领域	访问方式
VIVOS [Luong and Vu, 2016]	15	人工	朗读语音	Open
FOSD/FPTVIET [FPT技术研究院, 2020]	∼25	人工	朗读语音	Open
FLEURS（越南语） [Conneau等人, 2022]	∼12	人工	平行朗读	Open
VLSP2020 [VLSP组委会, 2020]	∼250	人工	新闻 + 自发	受限
VLSP2021 [VLSP组委会, 2021]	∼280	人工	通用	受限
VLSP2023 [VLSP组委会, 2023]	–	人工	通用	受限
VLSP2023-voting [VLSP组委会, 2023]	–	伪	通用	受限
CommonVoice（越南语） [Ardila 等人, 2020]	∼17	众包	众包	Open
VietMed [Le-Duc, 2024]	16	人工	医学	Open
VIET_BUD500 [Pham 等, 2024a]	∼500	伪	播客 / 多领域	Open
VietSpeech [Pham 等, 2024b]	∼1,100	伪	社交媒体	Open
Sub-GigaSpeech2-Vi [Yang 等人, 2024]	∼6,039	伪	多领域YouTube	Open
viVoice [Le 等人, 2024b]	∼1,017	伪	YouTube（文本转语音精选）	Open
Sub-PhoAudioBook [Vu 等人, 2025]	941	伪	有声书旁白	Open
VietSuperSpeech（本文）	**267.39**	**伪**	**闲聊、视频博客、海外侨民**	**Open**

麦克风和设备差异。 创作者使用各种各样的录音设备，从专业麦克风到手机摄像头，带来了与呼叫中心和移动聊天机器人部署相关的真实信道失真。

4.3 语言学特性

越南语是一种单音节、有声调的孤立语，具有六个词汇声调以及显著的方言音系变异 [Luong and Vu, 2016]。口语越南语还具有正式语料库中所没有的语域特定词汇、口语缩略形式和语码转换表达。VietSuperSpeech中的所有转写文本均保留完整的附加符号标记，包括声调符号，从而保留非正式语域在音系与词汇上的丰富性。

平均话语长度为266个字符，3–30秒的分段窗口与客服对话中典型的话轮长度高度匹配：单个客户话语很少超过30秒，且往往落在5–15秒区间。

5 Applica 应用：面向越南语人工智能系统的对话语音识别

VietSuperSpeec h 专为支持以下具有重大影响的应用领域而设计 .

5.1 语音聊天机器人系统

现代越南语语音聊天机器人——无论部署于银行业、电子商务、医疗保健还是公共服务领域——都必须能够准确识别不同说话者通过消费级设备说出的日常、口语化越南语。基于正式语料微调的自动语音识别系统，在处理非正式的查询表达、地区口音和口语词汇时，识别效果往往会明显下降。VietSuperSpeech 提供了一个有针对性的微调语料库，使模型能够接触到用户与

语音助手交谈时自然采用的非正式语体，从而提高首次识别准确率，减少需要反复澄清的情况。

5.2 客户支持与联络中心自动化

在越南快速发展的服务型经济中，对客户支持来电进行自动转写是一项价值很高的自然语言处理应用。此类情境中的客户语音自发性强，往往带有浓厚情绪，经常夹杂产品名称和服务术语与口语表达，并且经过电话编解码压缩传输。VietSuperSpeech 的声学和语言多样性——尤其是其中的海外侨民和非正式评论内容——与这一应用场景高度契合。机构可以使用 VietSuperSpeech，对通用越南语自动语音识别模型进行微调，使其适应非正式语体的客户对话，之后再利用专有呼叫中心音频进一步适配。

5.3 热线与交互式语音应答系统

部署在越南电话热线上的交互式语音应答（IVR）系统，必须服务广泛的来电者群体，他们的地区口音、语速以及对技术媒介通信的熟悉程度各不相同。VietSuperSpeech 中的海外侨民社区内容涵盖了可能混合地区口音或进行语码转换的说话者，而个人视频博客内容则呈现了技术使用经验较少者非正式、自主表达的语音风格。两者相结合，使 VietSuperSpeech 成为适用于交互式语音应答系统的自动语音识别组件的实用预训练或微调资源。

5.4 低资源领域自适应

对于缺乏专有越南语会话数据的从业者而言，VietSuperSpeech 可作为桥接数据集。将自动语音识别模型部署于面向客户的应用的机构，可以先使用 VietSuperSpeech 对通用越南语模型（例如，基于 Sub-GigaSpeech2-Vi [Yang 等人， 2024]或 VIET_BUD500 [Pham 等人，2024a] 与 VLSP 数据的组合训练的模型）进行微调，以实现会话领域适应；随后再使用有限的专有数据进行最终微调。这种多阶段方法与高资源自动语音识别领域自适应中的成熟做法一致 [Radford 等，2023, Gandhi 等人， 2023]。

5.5 与现有语料库互补使用

VietSuperSpeech 旨在补充而非取代现有的越南语自动语音识别数据集。我们建议在多语料训练流程中按以下方式将其结合使用：

- **正式／清晰语音覆盖**：VIVOS [Luong and Vu， 2016], FOSD [FPT技术研究院， 2020], Sub-PhoAudioBook [Vu 等人， 2025];
- **通用大规模覆盖**：Sub-GigaSpeech2-Vi [Yang 等人， 2024], VIET_BUD500 [Pham 等人，2024a], VietSpeech [Pham 等人，2024b];
- **特定领域覆盖**：VietMed [Le-Duc， 2024] （医学），VLSP2020/2021年/2023 [VLSP组委会, 2020, 2021, 2023] （新闻和通用）；
- **对话式覆盖范围（本研究）**：VietSuperSpeech。

6 局限性与伦理考量

伪标签噪声。与所有伪标注语料库一样 [Pham 等人，2024a，Yang 等，2024，Le 等人，2024b]，VietSuperSpeech 的转写文本可能包含系统性错误，尤其是在处理不流畅表达、重叠语音或罕见词汇时。教师模型（Zipformer-30M-RNNT-6000h [Nguyen,2025]）主要使用干净的越南语语音进行训练，偶尔可能会错误转写严重不流畅的表达或带有浓重口音的语音。

语域覆盖。 尽管 VietSuperSpeech 着重收录随意语音，但其来源频道受 YouTube 上可获取内容的限制，可能无法充分代表高度专业化的会话子语域（例如，在非常嘈杂的呼叫中心环境中，或由不熟悉技术的老年说话者所使用的越南语）。

人口统计与方言平衡。 说话人群体并未经过人口统计学控制。北部、中部和越南南部方言以及海外侨民变体的比例由来源频道决定，可能与实际部署的客户服务场景中的分布不成比例。

版权与隐私。 所有音频均从公开访问的 YouTube 内容中收集，并遵循非商业研究的合理使用原则 [Le-Duc，2024]。除音频内容本身外，我们未收集或存储任何个人身份信息。该数据集仅用于非商业研究。

7 数据可用性

VietSuperSpeech 可通过 HuggingFace Datasets 公开获取：

https://huggingface.co/datasets/thanhnew2001/VietSuperSpeech

该数据集以 HuggingFace Parquet格式分发，以便进行高效的流式处理和加载，音频以 16kHz单声道 WAV 格式存储。仓库页面提供了完整的文档，包括字段说明、加载代码示例和推荐的微调方案。

8 结论

我们介绍了 VietSuperSpeech，这是一个包含 52,023 条话语、时长达 267.39 小时的越南语自动语音识别数据集，经过精心整理，重点收录来自 YouTube 个人视频博客、非正式评论和海外越南裔社群频道的随意对话语音。现有越南语自动语音识别语料库系统性地忽视了非正式语体，而这些语料库主要涵盖朗读语音（VIVOS [Lu- ong and Vu，2016]，FOSD [FPT技术研究院，2020]，FLEURS [Conneau等人,2022]）、有声书（Sub-PhoAudioBook [Vu 等人，2025]）、播客（VIET_BUD500 [Pham et al., 2024a]）以及通用网络爬取语料库（Sub-GigaSpeech2-Vi [Yang 等人，2024]，VietSpeech [Pham et al., 2024b]）。VietSuperSpeech 填补了这一关键空缺，可供开发越南语聊天机器人、客户支持、呼叫中心和热线自动语音识别系统的从业者使用。我们开放发布该数据集，并鼓励将其与更广泛的越南语自动语音识别语料库生态系统结合使用，作为会话领域适应资源。

致谢

我们感谢 Sherpa-ONNX 的开发者和 k2-fsa 社区提供开源语音推理工具，也感谢 Thien HyNguyen 开发并公开发布 Zipformer-30M-RNNT-6000h 模型 [Nguyen，2025]，该模型为 VietSuperSpeech 的伪标签流程提供了基础。

参考文献

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. 和 Weber, G. Common Voice：一个大规模多语种语音语料库。载于第十二届语言资源与评测会议（*LREC*）论文集，第 4211–4215 页，2020年。

Baevski, A., Zhou, Y., Mohamed, A. 和 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（*NeurIPS*），第 33 卷，第 12449–12460 页，2020年。

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C. 和 Bapna, A. FLEURS：通用语音表征的少样本学习评测。载于 *IEEE* 口语语言技术研讨会（*SLT*），第 798–805 页。IEEE，2022年。arXiv:2205.12446。

FPT技术研究院。FPT 开放语音数据集（FOSD）。https://data.mendeley.com/datasets/k9sxg2twv4/4，2020年。

Gandhi, S., von Platen, P. 和 Rush, A. M. Distil-whisper：通过大规模伪标签实现稳健的知识蒸馏。*arXiv* 预印本 *arXiv:2311.00430*，2023年。

Graves, A. 基于循环神经网络的序列转导。*arXiv* 预印本 *arXiv:1211.3711*，2012年。

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. 和 Pang, R. Conformer：用于语音识别的卷积增强型 Transformer。载于*Interspeech* 论文集，第5036–5040页，2020年。

Le, T.-T., Nguyen, L. T., and Nguyen, D. Q. PhoWhisper：面向越南语的自动语音识别。载于*ICLR 2024* 第二届微型论文专场，2024。

Le-Duc, K. VietMed：面向医疗领域越南语自动语音识别的数据集与基准。载于*2024* 年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第 17365–17370 页，意大利都灵，2024。arXiv:2404.05659。

Le Phuoc Gia, T., Pham Minh, T., Nguyen Quoc, H., Nguyen Quoc, T., and Truong Hoang, V. viVoice：实现越南语多说话人语音合成。技术报告，胡志明市 FPT 大学，2024。https://huggingface.co/datasets/capleaf/viVoice。

Luong, H.-T. 和 Vu, H.-Q. 面向越南语语音识别系统的非专家 Kaldirecipe。收录于《全球语言服务基础设施第三届国际研讨会暨人类语言技术开放基础设施与分析框架第二届研讨会论文集（*WLSI/OIAF4HLT2016*）》，第 51–55 页，日本大阪，2016。

Nguyen, B. P. 等。多方言越南语：任务、数据集、基线模型与挑战。*arXiv* 预印本 *arXiv:2410.03458*，2024。Pham, A.、Tran, K. L.、Nguyen, L.、Cao, T. D.、Phan, P. 和 Nguyen, D. A. BUD500：综合性越南语自动语音识别数据集。技术报告，VietAI Research，2024。https:// github.com/quocanh34/Bud500。Pham, Q. N.、Pham, H. A. D. 和 Nguyen, V. T. VietSpeech：越南语社交语音数据集。技术报告，MMLab-UIT，2024。https://huggingface.co/datasets/NhutP/VietSpeech。Radford, A.、Kim, J. W.、Xu, T.、Brockman, G.、McLeavey, C. 和 Sutskever, I. 基于大规模弱监督的鲁棒语音识别。收录于《第 *40* 届国际机器学习会议论文集（*ICML*）》，第 28492–28518 页，2023。Shriberg, E. E. 言语不流利现象理论初探。博士论文，加州大学伯克利分校，1994年。VLSP组委会。VLSP 2020 自动语音识别挑战赛。https://vlsp. org.vn/vlsp2020/eval/asr，2020年。VLSP组委会。VLSP 2021 自动语音识别挑战赛。https://vlsp. org.vn/vlsp2021/eval/asr，2021年。VLSP组委会。VLSP 2023 自动语音识别挑战赛。https://vlsp. org.vn/vlsp2023/eval/asr，2023年。Vu, T.、Nguyen, L. T. 和 Nguyen, D. Q. 越南语零样本文本转语音。收录于《第 *63* 届计算语言学协会年会论文集（*ACL 2025*），2025年。arXiv:2506.01322。Ward, N. G. 《英语会话韵律模式》。剑桥大学出版社，2019年。Yang, Y.、Song, Z.、Zhuo, J.、Cui, M.、Li, J.、Yang, B.、Du, Y.、Ma, Z.、Liu, X.、Wang, Z.、Li, K.、Fan, S.、Yu, K.、Zhang, W.-Q.、Chen, G. 和 Chen, X. GigaSpeech 2：通过自动爬取、转写和精炼构建的不断发展的低资源语言大规模多领域自动语音识别语料库。*arXiv* 预印本 *arXiv:2406.11546*，2024。Nguyen, H. T. 越南语语音识别与情感识别综合流程。收录于《第 *11* 届越南语语言与语音处理国际研讨会论文集（*VLSP 2025*），2025年。模型可在 https://huggingface.co/hynt/ Zipformer-30M-RNNT-6000h 获取。Yao, Z.、Wu, L.、Yang, X.、Kang, W.、Kuang, F.、Yang, Y.、Jin, Z.、Lin, L. 和 Povey, D. Zipformer：一种更快、更出色的自动语音识别编码器。收录于《第 *12* 届国际学习表征会议论文集（*ICLR*），2024。arXiv:2310.11230。Zavaliagkos, G. 和 Colthurst, T. 利用未转写训练数据提升性能。收录于《*DARPA*广播新闻转写与理解研讨会》，第 301–305 页，1998年。