面向印尼语会话音频的 Pyannote 说话人日志流水线领域自适应

Muhammad Daffa'I Rafi Prasetyo[1], **Ramadhan Andika Putra**[1],**Zaidan Naufal Ilmi**[1],和 **Kurniawati Azizah**[II]计算机科学学院，印度尼西亚大学{muhammad.daffai, ramadhan.andika, zaidan.naufal}@ui.ac.id
kurniawati.azizah@cs.ui.ac.id

arXiv:2601.03684v1 [cs.SD] 7 Jan 2026

摘要

本研究提出一种面向印尼语会话音频的说话人分离领域自适应方法。我们采用神经文本转语音技术进行合成数据生成，以应对将以英语为中心的说话人分离流程适配至低资源语言所面临的挑战。我们在不同训练配置下进行了实验，使用了一个小型数据集（171 个样本）和一个包含 25 小时合成语音的大型数据集。结果表明，在 AMI 语料库上训练的基线 pyannote/segmentation-3.0 模型，以零样本方式应用于印尼语时，说话人分离错误率（DER）为 53.47%。领域自适应显著提升了性能：小型数据集模型将 DER 降至 34.31%（1 个训练轮次）和 34.81%（2 个训练轮次）。在 25 小时数据集上训练的模型性能最佳，DER 为 29.24%，较基线绝对提升 13.68%，同时保持 99.06% 的召回率和 87.14% 的 F1 分数。

索引词—说话人分离、领域自适应、Pyan-note、印度尼西亚语语音处理、合成数据

I. 引言

说话人分离通常被定义为确定"谁在何时讲话"的任务，在提升自动转录文本的可读性，以及为商务会议、呼叫中心和广播媒体等场景提供丰富的分析功能方面发挥着关键作用。近年来，基于神经网络的方法发生了范式转变，端到端神经说话人分离（EEND）和 pyannote.audio 等模块化流程 [1] 树立了新的性能基准。

尽管取得了诸多进展，不同语言的说话人分离模型性能仍存在显著差异。预训练模型主要针对高资源语言进行优化，尤其是英语，并使用 AMI 会议语料库等大型数据集 [2]。将这些模型应用于印度尼西亚语会话音频时，由于语言特征、声调特征和声学环境各不相同，模型会受到领域不匹配的影响。由于公开可用的印度尼西亚语会话数据集数量稀少，且缺少标注，因此在说话人分离领域中，印度尼西亚语被归为低资源语言 [3]。

本研究探讨如何将稳健的英语中心说话人分离系统适配到印度尼西亚语领域。我们采用领域自适应 [4]策略，假设说话人区分的低级特征具有可迁移性，而分割模型则需要进行适配，以处理印度尼西亚语语音特有的时间动态特征和语音学特性。为避免训练数据不足，我们采用数据合成策略，利用神经文本转语音（TTS）[5] 生成带有真实标注的逼真会话音频。

本研究的贡献有三方面：

- 开发了一套合成数据生成流程，用于创建带标注的印度尼西亚语会话语音。
- 应用领域自适应技术对 pyannote 分割模型进行微调，有效地将知识从 AMI 语料库迁移到印度尼西亚语情境中。
- 通过实证评估表明，与基线模型相比，自适应后的模型实现了更低的说话人分离错误率（DER） [6]和更高的召回率。

II. 方法论

本研究为语音说话人分离制定了一种领域自适应策略。领域自适应是一种迁移学习技术，利用源领域的知识来提升模型在数据分布不同的目标领域上的表现 [4]。具体而言，我们构建了一套流程，以弥合从英语（源语言）到印度尼西亚语（目标语言）的语言差异。如图 1 所示，实验工作流程分为四个阶段：合成数据生成、模型训练、推理和评估。

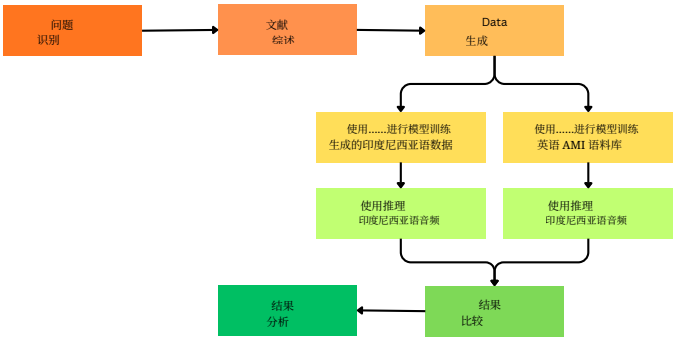

图 1. 研究方法流程

A. 合成数据生成

要实现对印度尼西亚语的领域自适应，必须具备目标语言的会话数据。此要求不仅涉及原始音频，还包括精确的时间标注，具体而言，即富转录时间标记（RTTM）格式的数据。这类标注数据对于语音说话人分离任务至关重要，因为它提供了必要的真实标注，标明“谁在何时说话”。然而，高质量的印度尼西亚语会话数据集标注仍然十分稀缺；在语音处理领域，印度尼西亚语通常被归类为低资源语言 [3]。

为解决数据稀缺问题，本研究采用了基于文本转语音（TTS）技术的数据合成策略。合成流程利用 edge-tts 库 [5]生成高质量的神经语音波形。为构建会话结构，我们使用 Pydub 框架 [7], 处理并混合单段语音，从而精确控制时间安排和重叠逻辑。借助该系统，我们生成了 25 小时的印尼语会话音频样本，并配有相应的 RTTM 文件；这些样本经过专门设计，包含重叠语音片段。

B. 基线系统

为建立性能基准，本研究采用 pyannote/segmentation-3.0 流程进行语音说话人分离 [1]。该基线模型使用 AMI 会议语料库 [2]，这是一个大型多模态数据集，包含约 100 小时的会议录音。该语料库收录了英语中的自然多说话人对话，呈现了真实会议环境中的复杂动态。

在本研究中，基于 AMI 语料库训练的模型作为源领域代表。该模型在零样本设置下进行评估，未使用印度尼西亚语数据进行微调。通过将这一基线模型与使用印度尼西亚语语音训练的领域适配模型进行比较，我们可以定量评估由语言领域偏移造成的性能差距，并评估纳入特定语言训练数据的效果。

C. 领域自适应

为了解决上一节所述的领域偏移，我们采用了一种迁移学习策略，重点适配说话人分离流程中的分割模块。我们使用了与基线系统相同的预训练 pyannote/segmentation-3.0 架构。不过，我们没有从头开始训练模型（这需要大量标注数据），而是使用从 AMI 语料库中学到的参数初始化模型权重。这样的初始化使模型能够保留从源领域中学到的通用说话人区分能力，同时为适配印度尼西亚语的特定声学特征做好准备。

适配过程采用监督微调，并使用按照自定义协议组织的合成数据集。通过最小化模型在该目标协议上的损失，模型能够有效调整其内部表示，以识别印度尼西亚语的语音学特征和韵律，从而弥合源领域与目标领域之间因语言不匹配而造成的性能差距。

D. 评估指标

客观比较结果是本研究的基本环节。为确保能够量化评估模型性能，需要采用广受认可的标准指标。因此，我们采用说话人日志错误率（DER）作为主要评估基准。DER 定义为未正确归属给相应说话人的输入语音时长占比，其中汇总了语音活动检测和说话人聚类产生的错误 [6]。该指标的数学计算公式如下 [6]:

$$DER = \frac{T_{FA} + T_{MISS} + T_{CONF}}{T_{TOTAL}} \times 100\% \qquad (1)$$

其中:

- T_{FA} (虚警): 在静音期间检测到语音。
- T_{MISS} (漏检): 未检测到语音。
- T_{CONF} (混淆): 语音被分配给了错误的说话人。
- T_{TOTAL}: 参考标注中的语音总时长。

说话人分离错误率越低，表示说话人分离性能越好。

III. 实验

为验证所提出的领域自适应策略，我们进行了一系列实验，将基于 AMI 训练的基线模型与适配印度尼西亚语的模型进行比较。本节详细介绍了为确保实验可复现而采用的实现框架、数据配置和训练超参数。

A. 实现框架

实验流程使用开源工具包 pyannote.audio（版本 3.1）实现， [1],该工具包是我们说话人分离任务的基础。训练过程借助 pytorch-lightning 管理优化循环和 GPU 加速。所有实验均在配备 NVIDIA T4 图形处理器的计算环境中进行，以处理分割模型的计算负载。

B. 数据配置与协议

实验中的关键步骤是定义一个自定义实验协议，将合成数据与训练流程衔接起来。我们在系统的数据库注册配置（database.yml）中定义了一个名为 DebateIndonesianLarge 的协议。

该协议严格将生成的合成数据集划分为三个子集，以防止数据泄漏：

- **训练集：** 用于对模型权重进行监督微调。
- **开发集：** 用于验证和超参数调优。
- **测试集：** 仅用于最终推理和说话人分离错误率计算。

通过注册此协议，我们确保模型能够在训练轮次中系统地加载音频和 RTTM 真实标注。

C. 训练超参数

在微调阶段，我们实例化了框架提供的分割任务。我们采用了专门针对捕捉对话动态而优化的超参数：

- **分块时长：** 2.0 秒音频片段
- **批次大小：** 每批 16 个音频块
- **训练轮数：** 我们试验了 1 和 2 个训练轮次
- **GPU 加速：** 支持 CUDA 的 NVIDIA T4

模型经过训练，直至验证损失趋于稳定，从而确保领域自适应过程在最终评估前已收敛。

D. 数据集配置

我们采用两种数据集配置进行了实验，以评估训练数据量的影响：

- **小型数据集：** 171 个合成印尼语对话音频样本
- **大型数据集（25 小时）：** 一个扩展数据集，包含约 25 小时的合成对话式印度尼西亚语语音，用于评估增加数据量是否能够提升模型泛化能力

IV. 结果与分析

为评估领域自适应流水线的有效性，我们使用标准检测指标分析其定量性能。通过将所提出的模型与预训练基线进行比较，本评估建立了一个基准。

A. 定量性能

该评估比较了三种实验配置：AMI 基线（以零样本方式应用预训练的 pyannote/segmentation-3.0 模型）、小规模自适应模型（在 2 小时合成印度尼西亚语数据上进行微调）以及大规模自适应模型（在 25 小时合成数据上进行微调）。

表 I 性能比较：AMI 基线与印度尼西亚语适配模型（2 小时与 25 小时）

指标	AMI 基线	印尼语适配版（2 小时）	印尼语适配版（25 小时）
精确率	68.18%	74.18%	**77.78%**
召回率	87.23%	**100.00%**	99.06%
F1 分数	76.54%	85.17%	**87.14%**
DER	53.47%	34.81%	**29.24%**

如表 I 所总结，所提出的领域自适应策略相比基线取得了显著改进。仅引入 2 小时的合成数据，就将说话人分离错误率（DER）从 53.47%（基线）大幅降至 34.%，同时实现了完美的灵敏度，召回率达到 100.00%。这表明，即使少量领域内数据也能有效弥合最初的语言差距。

此外，将合成数据集扩展至 25 小时优化了模型的判别能力。2 小时模型的灵敏度很高，而 25 小时模型的精确率更高，从 74.18% 提升至 77.78%。假阳性减少，使模型取得了最佳的整体表现：说话人分离错误率最低，为 29.24%；F1 分数最高，为 87.14%。这一趋势表明，规模更大的合成数据集对于提升模型区分真实语音重叠与噪声的能力至关重要。

B. 错误分布分析

对混淆矩阵的比较揭示了适配过程如何改变模型的错误特征。如图2所示，AMI 基线难以应对领域偏移，表现出最高的漏检率，未能识别出 12.77% 的重叠语音片段（召回率：87.23%）。此外，其精确率较低（68.18%），表明模型容易产生误报，将单人语音错误地标记为重叠语音。

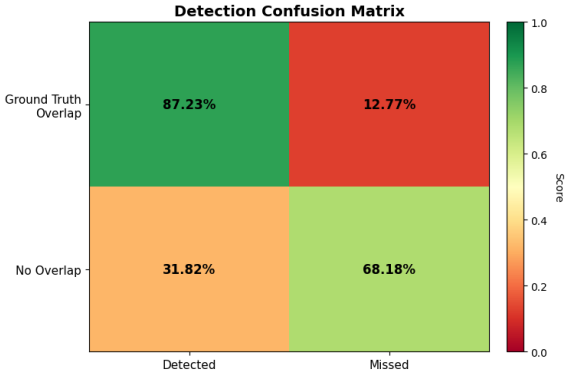

图2. AMI 基线的检测混淆矩阵，显示漏检率较高（12.77%），且假阳性率较高。

相比之下，2小时印尼语适配模型采取了更激进的检测策略。如性能指标所示，该模型的召回率达到 100.00%，成功消除了漏检错误。然而，这种灵敏度是以精确率（74.18%）为代价的；如图3所示，该模型优先捕捉所有可能的重叠语音。

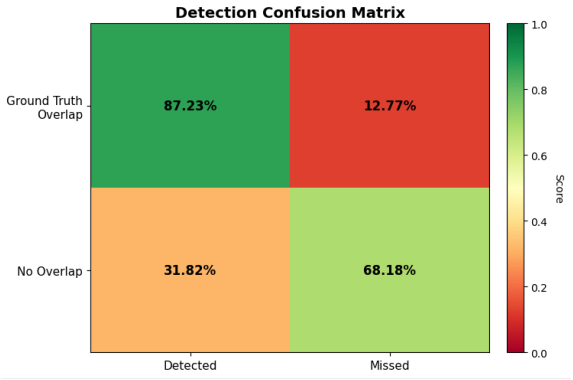

图3. 2小时适配模型的检测混淆矩阵，该模型实现了完美的灵敏度（召回率为 100%）。

25Hours 印度尼西亚语适配模型展现出最均衡、最稳健的性能。如图4所示，该模型成功保持了近乎完美的灵敏度，仅漏检了 0.94% 的重叠语音实例（召回率：99.06%）。至关重要的是，增加训练数据使模型能够比 2 小时模型更有效地区分真实的重叠语音与噪声。这一点可从假阳性率的下降看出（由“无重叠—检测到”象限表示），该比例降至 22.22%，对应于最高的精确率 77.78%。

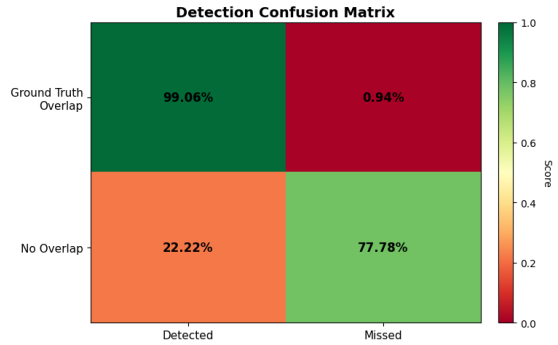

图4. 25小时适配模型的检测混淆矩阵。与基线相比，该模型在最大限度减少漏检（0.94%）的同时，有效降低了误报。

C. 领域自适应的有效性

实验结果验证了对印尼语会话音频进行领域自适应的必要性。基线模型的说话人分离错误率（DER）高达 53.47%，凸显了源领域（英语会议）与目标领域（印度尼西亚辩论）之间严重的语言和声学不匹配。虽然基线模型能够检测出相当一部分重叠语音，但仍漏掉了 12.77% 的关键片段，且精确率较低。

借助印尼语合成数据集，自适应后的模型成功弥合了这一差距。经过 2 小时的自适应，灵敏度问题便得到纠正，几乎不再出现漏检。此外，将数据扩展至 25 小时证明，该流程不仅可以针对灵敏度进行优化，也能针对精确率进行优化。这一全面策略使说话人分离错误率绝对降低了 24.23%（从 53.47% 降至 29.24%），证实了使用合成数据进行迁移学习对低资源语言非常有效。

五、结论

本研究成功证明了领域自适应在印尼语会话音频说话人分离任务中的有效性。我们借助神经文本转语音技术生成的合成数据，将预训练的 pyannote/segmentation-3.0 模型自适应至印尼语语言领域。

我们的实验得出了三个关键结论：

1) **领域自适应至关重要：**基线模型在印度尼西亚语上的零样本性能（说话人分离错误率为 53.47%）证实，英语与印度尼西亚语的语音模式之间存在显著的领域不匹配。2) **训练数据量很重要：**小型数据集（171 个样本）将说话人分离错误率降低至约 34%，而包含 25 小时数据的数据集则将其进一步降至 29.24%，这表明更大的训练数据集能够提升模型泛化能力。3) **确定了最佳配置：**使用 25 小时合成印度尼西亚语语音训练 2 个轮次的模型，在精确率（77.78%）、召回率（99.06%）和整体性能（F1 分数：87.14%，说话人分离错误率：29.24%）之间实现了最佳平衡。

未来的工作可以探索使用真实场景的印度尼西亚语会话数据进行微调，将重叠语音检测与说话人嵌入提取相结合的多任务学习方法，以及将该方法应用于其他低资源语言。

VI. 建议

基于本研究的发现，我们提出以下未来工作建议：

1) **纳入真实场景数据：**将合成数据与真实的印度尼西亚语会话录音相结合，可以进一步提升模型鲁棒性及其在真实场景中的泛化能力。2) **扩充训练数据：**鉴于使用 25 小时数据集所观察到的显著改进，生成更大的合成数据集（50+小时）可能会进一步提升性能。3) **超参数优化：**在特定领域验证集上对流水线参数进行微调，可以针对具体应用优化精确率-召回率权衡。

参考文献

[1] H. Bredin 等人，《Pyannote.audio：用于说话人分离的神经网络构件》，载于电气与电子工程师协会声学、语音与信号处理国际会议论文集（*ICASSP*），2020年，第7124–7128页。[2] AMI联盟，《AMI会议语料库》，爱丁堡大学。 [在线]。获取地址：https://groups.inf.ed.ac.uk/ami/corpus/。[访问日期：11月 19, 2025]。[3] A. Fain 和 A. Lay，《面向低资源印度尼西亚语的端到端语音识别模型》，载于国际计算机科学与信息系统进展会议论文集（*ICACSIS*），2020年。[4] F. Zhuang 等人，《迁移学习综合综述》，电气与电子工程师协会会刊，第109卷，第1期，第43–76页，2021年1月。[5] S. Shize，《edge-tts：用于调用 Microsoft Edge 在线文本转语音服务的 Python 库》，GitHub 仓库，2023年。 [在线]。获取地址：https://github.com/rany2/edge-tts。[6] J. G. Fiscus、J. Ajot、M. Michel 和 J. S. Garofolo，《〈2006年春季丰富转录（RT-06S）评测计划〉》，美国国家标准与技术研究院，美国，技术报告，2006年。 [在线]。获取地址：https://www.nist.gov/itl/iad/mig/rich-transcription-spring-2006-evaluation-plan

[7] J. Robert，《Pydub：通过简单易用的高级接口处理音频》，GitHub 仓库，2018年。 [在线]。获取地址：https://github.com/jiaaro/pydub。