arXiv:2302.05110v1 [eess.AS] 10 Feb 2023

采用领域多样化与泛化的跨语料库口语语言识别

Spandan Dey[a], *, Md Sahidullah[b], **, Goutam Saha[a]

[a]印度理工学院克勒格布尔分校电子与电气通信工程系，印度克勒格布尔，邮编 721302[b]洛林大学、法国国家科学研究中心、法国国家信息与自动化研究院、洛林计算机科学研究实验室，法国南锡，邮编 F-54000

摘要

本文针对低资源口语语言识别（LID）问题中的跨语料库泛化问题展开研究。我们以印度语言识别为背景开展实验，发现由于语料库相关的非语言偏差，跨语料库泛化能力明显较差。本文的贡献有两方面。首先，我们提出领域多样化方法，利用不同的音频数据增强方法扩充有限的训练数据。随后，我们提出最大多样性感知级联增强的概念，并优化增强折叠因子，以有效实现训练数据多样化。其次，我们提出将增强方法视为伪域的域泛化思路。为此，我们研究了域不变方法和域感知方法。我们的语言识别系统采用当前最先进的增强型通道注意力、传播与聚合时延神经网络（ECAPA-TDNN）架构。我们使用印度语言识别研究中广泛采用的三个语料库开展了大量实验。此外，我们还在野外采集的 VoxLingua107 语料库印度语子集上，对所提方法进行了最终盲测评估。实验表明，与常用的简单增强方法相比，所提出的领域多样化方法更有前景。研究还表明，域泛化比领域多样化更为有效。我们还发现，对于同语料库语言识别，域感知学习表现更好；而对于跨语料库泛化，域不变学习更为适用。与基础 ECAPA-TDNN 相比，其所提出的域不变扩展将跨语料库等错误率最高降低了 5.23%。相比之下，所提出的域感知扩展也提升了同语料库测试场景下的性能。

关键词：语言识别（LID）、跨语料库评估、音频增强、域泛化、域对抗训练、多任务学习。

*通讯作者。

**Md Sahidullah 目前是一名独立研究员和顾问。电子邮箱： sd21@iitkgp.ac.in (Spandan Dey)、 md.sahidullah@inria.fr (Md Sahidullah)、gsaha@ece.iitkgp.ac.in (Goutam Saha)

提交至《计算机语音与语言》的预印本 2023年2月13日

1. 引言

在语音助手、智能家居和众多其他创新设备盛行的时代，基于语音的人机交互（HCI）正逐渐融入我们的日常生活。全球范围内，依赖现代语音应用的用户数量将持续增长，这些应用包括自动语音识别（ASR）、说话人识别与验证、语音合成和语音情感识别（SER） [14]。为便于使用，语音应用应能在多种输入语言下高效运行。通常会使用前端语言识别（LID）模块来实现多语言适用性 [69]。LID 系统根据输入预测语言，后续的语音应用则据此调整运行模式 [42]。

为了在现实世界中高效部署语音应用，提高前端语言识别（LID）模块的泛化能力至关重要。对于语言识别系统，泛化能力强的分类器应能抵御多种非语言因素的影响，例如说话人身份、性别、年龄、方言和口音，以及信道和背景环境不匹配所造成的影响 [17]。我们可以假设，随着数据采集环境的多样性增加，大型语音语料库中的非语言因素也会更加多样化。此类语料库或可通过减少与未知测试话语之间的不匹配来提高泛化能力 [83]。因此，拥有规模大且多样化的语料库，对于开发稳健的语言识别系统至关重要。

通过 NIST LRE [31, 74] 和 OLR 挑战赛 [44, 84], 等项目收集的语音语料库，是经过仔细核验真实标签的大规模多语言语音语料库。这些语料库被广泛用于采用深度神经网络（DNN）架构开发和评估语言识别算法 [30, 49, 57], ，在过去十年中极大推动了语言识别研究。然而，由于缺乏标准的大规模语料库，世界上许多广泛使用的语言仍属于低资源语言。针对这些低资源语言的语言识别研究通常使用独立创建的小规模内部语料库开展，而这类语料库的非语言变体有限 [18]。因此，使用此类小型语料库训练最先进的基于 DNN 的语言识别系统容易导致过拟合。为解决这一问题，研究人员尝试了多种方法，例如采用数据增强、特征选择、正则化器、交叉验证或模型集成 [17, 70, 78]。然而，如何让低资源语言的语言识别系统实现泛化，仍是关键挑战 [17]。

本研究旨在提升跨语料库框架下低资源语言识别系统的泛化能力。在语言识别领域的文献中，大多数研究通过训练集训练系统，并在同一语料库专用的评估集上对其进行评估。然而，对于低资源语料库而言，数据集内的数据多样性可能有限。因此，即使在自身评估集上取得了不错的性能，其在不同数据上的评估中能否泛化仍值得探究。因此，我们通过跨语料库评估来考察语言识别系统的泛化能力。对于每个语料库，我们分别训练语言识别模型，并使用所有可用语料库的测试子集对其进行评估。我们针对不同低资源语言识别语料库开展的初步研究 [17] 发现，跨语料库语言识别性能明显低于同语料库性能。在该研究中，我们采用了基于梅尔频率倒谱系数（MFCC）特征的基础 x 向量时延神经网络（TDNN）架构 [82] 。本研究采用当前最先进的 TDNN 增强型通道注意力、传播与聚合架构（ECAPA-TDNN） [16] ，并在此基础上进一步改进，以提升泛化能力。

在本研究中，我们的核心重点是通过缩小

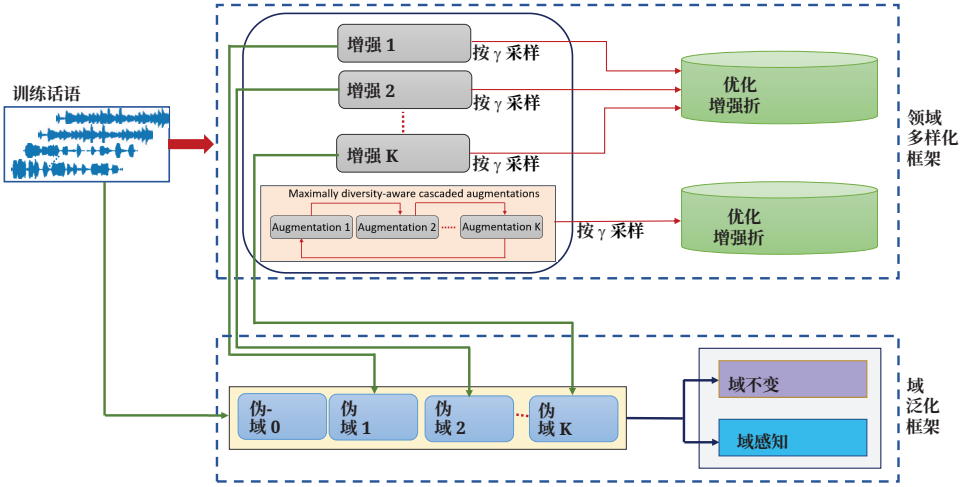

图1. 提升语言识别跨语料库泛化能力的拟议框架概览。此处，γ 表示增强扩增倍数（详见第5.3节），K表示信号级增强类别的总数。

同语料库评估中的性能失配，来解决跨语料库泛化问题。我们提出了新颖的领域多样化策略，并进一步引入域泛化（DG）概念 [10]，以此阐述该领域的跨语料库语言识别问题。解决跨语料库性能问题的一种可能方法是采用领域自适应（DA） [63, 100] 方法。在域自适应框架中，训练语料库被用作源域，并且我们可以获取来自评估条件的有限标注或未标注数据，将其称为目标域。同样，在跨语料库场景中，一个语料库的训练集可以视为源域，而其余语料库中的训练集话语则可以视为各自独立的目标域。然而，在某些实际场景中，无法直接应用域自适应方法，原因有二。首先，在训练语言识别系统时，我们可能完全无法获取特定目标域的音频数据。其次，即使我们拥有一些目标域数据（在我们的问题中，这些数据可以来自训练语料库以外的其他语料库），应用域自适应也只能提升模型在该特定目标语料库上的泛化能力，而对未见条件的泛化仍是一个关键问题。因此，我们在本研究中采用域泛化的概念。这是一种较新且实用的解决方案，无需特定目标域数据 [103]。相反，在域泛化框架中，我们从源域创建多个伪目标域，并训练模型以提升其在多个伪域上的泛化能力。我们采用多种领域多样化方法创建不同的伪域，随后利用这些伪域构建多种域泛化方法。我们通过扩展当前最先进的ECAPA-TDNN架构，提出域不变和域感知的域泛化方法，以学习语言判别知识。

图1概述了我们为提升语言识别系统泛化能力而提出的解决方案。图1展示了我们探索的两种主要方法：领域多样化和域泛化。

本文的主要贡献总结如下：

- 我们探索了多种音频数据增强方法，包括常用方法以及几种提高数据多样性的新方法。对不同

增强技术的全面研究也有助于其他基于语音的分类任务。

- 为了在原始数据量与增强数据量之间保持平衡，我们系统地研究了增强折叠因子对基于TDNN的语言识别分类器泛化能力的影响。该分析还从计算复杂度和泛化性能两方面给出了最优增强折叠因子。

- 我们提出最大多样性感知级联增强，以进一步提高低资源语料库的多样性。该方法基于最大均值差异的域间多样性度量生成更逼真的增强音频数据，同时保留口语的感知特性和可懂度。

- 我们将不同的域多样化音频类别视为多个伪域，并引入域泛化的概念，以提升跨语料库性能。我们通过引入两种方向相反但功能相似、适用于未知目标域的域泛化方法，对当前最先进的ECAPA-TDNN架构进行了扩展。第一种是域不变方法，在这种方法中，语言识别模型经过训练后，会在识别不同伪域的任务中产生混淆。在域不变方法中，我们探索了对抗训练和最大均值差异（MMD）方法。第二种是域感知方法；在学习语言判别线索的同时，我们训练语言识别模型识别不同的伪域。在域感知方法中，我们探索了将伪域分类作为附加任务的多任务学习。

- 我们进一步拓展了传统的基于最大均值差异的域泛化方法，提出了多域最大均值差异（MD-MMD）算法。该方法并非将所有伪域视为单一目标域，而是计算由多样性加权的多域最大均值差异损失，并联合优化这些MD-MMD损失。

- 我们使用三个独立收集的印度语言语言识别语料库，对同语料库和跨语料库语言识别性能进行了系统的基准测试与分析。最后，我们在第四个未见语料库上对这些算法进行了盲测评估。

本文其余部分安排如下：第2节正式介绍跨语料库评估范式。第3节介绍本文所用的语料库和实验方法。第4节讨论旨在提升跨语料库性能的域多样化技术。第5节则从域泛化方法的角度探讨跨语料库问题。第6节展示实验结果。第7节进一步分析跨语料库问题，包括在未见过的真实场景语料库上进行的盲测评估。最后，第8节总结全文，并提出建议。

2. 问题定义与相关工作

本节正式介绍了我们的跨语料库评估协议，并讨论了它在各种基于语音的分类任务中的应用。

2.1. 跨语料库评估的表述

在跨语料库评估范式中，模型在一个语料库上训练，并在多个语料库上测试。如果测试样本取自训练语料库中留出的部分，则称为同语料库评估；如果测试样本取自其他语料库，则称为跨语料库评估。接下来，我们将给出跨语料库评估的正式定义，并介绍相关观点与限制条件。

定义：考虑一个包含 C 个语料库的跨语料库实验设置。第 i 个语料库的数据表示为 $\{(\mathrm{X}_k^i, \mathrm{y}_k^i)\}^{N^i}{}_{k=1}$，其中 $\mathrm{i} \in [1,\mathrm{C}]$，$\mathrm{N}^i$ 表示语料库 i 中的样本数。X_k 是第 i 个语料库中第 k 个话语的特征，y_k是对应的标签。

对于每个语料库 i，特征空间 $\mathcal{X}^i = [\mathrm{X}_1^i, \mathrm{X}_2^i, \cdots, \mathrm{X}_{N_i}^i]$ 被划分为互不相交的训练部分和评估部分 $\mathcal{X}_T^i$ 和 $\mathcal{X}_E^i$。相应的标签空间 $\mathcal{Y}^i = [\mathrm{y}_1^i, \mathrm{y}_2^i, \cdots, \mathrm{y}_{N_i}^i] \in \{1, 2, \cdots, \mathrm{L}\}$（L 表示目标类别总数）也相应地划分为 $\mathcal{Y}_T^i$ 和 $\mathcal{Y}_E^i$。

考虑映射函数 $\mathrm{f}(\mathrm{X}; \mathrm{W}) : \mathrm{X} \longrightarrow \mathrm{y}$，其中 W 表示映射函数中可学习的参数。与 f 对应的多类分类损失函数记为 $\mathcal{L}(\mathrm{y}, \mathrm{f}(\mathrm{X}; \mathrm{W}))$，训练过程中需要将其最小化。

设模型使用第 i 个语料库的训练集 $(\mathcal{X}_T^i, \mathcal{Y}_T^i)$ 进行训练，并使用第 j 个语料库的评估集 $\mathcal{X}^j{}_E$ 进行测试。对于测试话语，该训练-测试对的后验预测概率向量记为 $\mathrm{p}\left(\mathcal{Y}_{jE} \mid \mathcal{X}^j{}_E;\ \mathrm{W}^i\right)$，其中 W^i 表示在 $\mathcal{X}_T^i$ 上训练得到的学习参数。当 $\mathrm{j} = \mathrm{i}$ 时，评估遵循同语料库配置。我们重点关注跨语料库条件，即 j, i 时的情况。随后，我们考虑性能指标计算函数 g，该函数以真实类别标签 $(\mathcal{Y})$ 和预测标签（$\mathrm{p}(\mathcal{Y} \mid \mathcal{X}; \mathrm{W})$）为输入，并计算一个表示分类性能的标量值。

设 $(\mathcal{X}_T^i, \mathcal{Y}_T^i)$ 用于训练一个独立的语言识别模型（f_i），该模型使用语料库 i 进行训练。跨语料库评估的主要目标是评估并最小化 f_i 分别在 $\mathcal{X}_E^i$ 和 $\mathcal{X}^j{}_E$ 上进行评估时的性能失配。跨语料库性能失配（对于语料库 i 和 j）定义为:

$$D_g(i, j) = \left| g\left(\mathcal{Y}_E^i, p(\mathcal{Y}_E^i \mid \mathcal{X}_E^i; \mathbf{W}^i)\right) - g\left(\mathcal{Y}_E^j, p(\mathcal{Y}_E^j \mid \mathcal{X}_E^j; \mathbf{W}^i)\right) \right| \tag{1}$$

本研究旨在最小化性能失配（$\mathrm{D}_g(\mathrm{i}, \mathrm{j}), \forall\ \{\mathrm{i}, \mathrm{j}\} \in [1,\mathrm{C}]$），同时不降低同语料库性能。我们可以通过估计语言识别模型的参数（W^i），使最终满足以下条件来实现这一目标。

$$\min_{\mathbf{W}^i \mid \mathcal{X}_T^i} \mathcal{L}\left(\mathcal{Y}_E^i, f(\mathcal{X}_E^i; \mathbf{W}^i)\right) \ \text{s.t}\ D_g(i, j) \to 0 \tag{2}$$

需要确保公式 2 满足 $\forall\ \{\mathrm{i}, \mathrm{j}\} \ni \mathrm{j}, \mathrm{i}$。

视角与约束：将训练好的独立模型部署到现实世界时，测试样本可能来自任何未见域。因此，对于这类情况，针对特定目标域进行泛化可能并非最高效的解决方案。我们的目标是提升模型在未见域上的泛化能力，并以多个语料库上的评估表现作为参考。因此，除训练语料库外，其余语料库中的话语均不作为目标域。在我们的评估框架中，考虑到语言识别系统在现实世界应用中面临的挑战，我们采用以下约束。

- 独立模型 (f_i) 仅使用单个语料库的训练数据 ($\mathcal{X}_{iT}$, $\mathcal{Y}_{iT}$) 进行训练，其中 $i \in [1, C]$。在系统开发过程中，不使用其他语料库训练样本中的话语 $\mathcal{X}_{jT}$ $\ni j, i$。

- 为改善训练语料库 i 与评估语料库 j 之间的跨语料库性能失配，我们禁止使用以语料库 j 训练的语言识别模型在同语料库评估中的反馈。这是为了确保泛化不会偏向语料库 j。

- 为了公平比较性能，我们采用具有固定语言集合的闭集语言识别问题。所有语料库共享相同的标签空间： $\mathcal{Y} = \{T\mathcal{Y}_i\}_{Ci=1}$。

可以将跨语料库研究视为机器学习文献中常用的域自适应（DA）方法的一种应用 [63, 100]。根据公式 2，在跨语料库研究的域自适应表述中，语料库 i 和 j 分别可视为源域和目标域。然而，我们的跨语料库泛化目标与域自适应方法的目标存在根本差异。域自适应适用于目标数据已预先知晓的场景。然而，我们关注的是在现实世界部署期间提升语言识别的泛化能力；在此类场景中，测试数据未知，且可能来自多种非语言域。对于这类域外泛化，我们的跨语料库评估任务表述更为恰当。与域自适应不同，我们不使用语料库 j 中的话语子集作为目标域数据来计算 W^i。相反，我们遵循域泛化（DG）[103] （其与域自适应在根本上相似）：不使用特定目标域，而是采用若干领域多样化方法，从源语料库 i 创建多个伪域。随后使用伪域数据计算 W^i。

2.2. 跨语料库评估相关工作

迄今为止，大多数跨语料库评估研究都针对计算机视觉应用开展，包括目标检测 [39, 15], 面部表情识别 [104], 深度伪造检测器 [61], 行人重识别 [94]。在语音处理应用中，跨语料库评估常用于评估鲁棒性，尤其是情感识别和说话人识别反欺骗等任务；由于数据收集受到限制，这些任务的数据集差异有限。

在语音情感识别（SER）领域， [77] 中报告的研究使用六个不同的语料库，通过说话人级特征归一化开展了跨语料库 SER。除非测试数据与训练数据具有相似的背景，否则跨语料库 SER 的性能会大幅下降。在 [102], 中，作者使用了六个情感语音语料库，并将无标注情感话语加入聚合训练数据，从而提升了跨语料库性能。其他一些 SER 研究也探讨了跨语料库评估 [91, 90, 86, 79, 26]。

在说话人识别反欺骗领域， [67] 中的研究使用两个语料库 ASVspoof2015 和 BTAS2016 开展了跨语料库实验，涵盖常见的欺骗攻击、语音合成和语音转换。该研究表明，在不同数据条件下，模型的泛化能力较差。 [40], 中的作者使用了两个反欺骗语料库 ASVspoof 和 AVspoof，并通过分数融合结合多个系统，提高了跨语料库泛化能力。另一项近期研究使用七个反欺骗语料库和多种分类器，展示了不同程度的泛化能力 [13]。该研究还分析了跨语料库性能与不同信号级描述符之间的相关性。

跨语料库评估也被用于检测语音中是否存在构音障碍 [29],、音乐情绪预测 [35], 、多模态语音活动检测（VAD） [5], 、语音增强 [64], 和语音识别 [88]。在所有这些研究中，作者均指出，由于数据条件存在较大差异，模型的泛化能力较差。

与上述讨论的其他语音相关任务相比，语言识别研究中采用跨语料库评估的研究较少。少数研究报告了跨语料库实验结果，尽管论文中并未明确提及这一点 [2, 23, 45, 46, 47, 89]。[89] 中的研究使用 VoxLingua107 的子集训练语言识别系统，并在 KALAKA-3 和 LRE07 上进行评估。他们评估了不同统计模型和神经网络模型在跨语料库语言识别中的鲁棒性。然而，他们并未进行额外的改进来提升跨语料库泛化能力。另一项近期研究 [23] 提出了一种用于语言识别的判别式层次模型，并使用 NIST LRE、RATS、NIST SRE、VoxLingua107、Panarabic、CALLHOME 和 CALLFRIEND 数据库的组合进行训练，使用 LRE15、LRE17、LASRS、KALAKA 和 Crowdsource（CROWD）数据库的组合进行评估。这项工作的目标是开发一种判别式概率线性判别分析（PLDA）框架，以作为常用生成式建模方法的更佳替代方案。 [45] 中的研究在语言识别中采用多任务学习融入层次化音系组合信息。他们分别在 NIST LRE17 和 AP17-OLR 数据库上评估了所提算法。LRE17 评估话语取自窄带 MLS14 和宽带 VAST 数据库，而这些数据库并未用于 LRE17 的训练和开发。 [46],中，作者分别使用 NIST LRE17 和 AP17-OLR 语料库，并通过压缩激励（SE）和线性瓶颈模块，消除基于预训练多语言 wav2vec2.0 XLSR 模型的语音表征中的冗余。 [2], 中，作者为 OLR21 挑战赛的无约束语言识别任务使用了自定义训练集。他们将 OLR21 的训练语句与 Mozilla Common Voice [59] 数据结合起来。随后，作者使用预训练的 wav2vec 2.0 模型，并利用自定义训练集对其进行微调。

上述大多数研究通过混合多个语料库的数据，违反了跨语料库评估标准，而且其主要目标并未明确聚焦于减少跨语料库性能失配。此外，这些研究主要关注语言资源丰富的主要语言。我们目前的工作致力于使用独立开发、资源有限的数据库，提升跨语料库语言识别性能。

3. 实验设置

本节介绍我们的实验设置，包括语音语料库、特征提取和神经网络架构的详细信息。

3.1. 语料库描述

我们选取了印度语言识别文献中最广泛使用的三个语料库：IIITH-ILSC（IIITH） [92], LDC 2017S14（LDC） [38], 和 IITKGP-MLILSC（KGP） [51]。在实验中，我们选用了三个语料库共有的五种语言：孟加拉语、印地语、旁遮普语、泰米尔语和乌尔都语。表 1 总结了这三个语料库的元数据。

表1中的元数据对比列出了这三个语料库之间的异同。语音时长表明，与主要语言的语言识别语料库相比，这些都是低资源语料库。其中，KGP规模最小，只有四

表 1：实验中使用的三个语音语料库的元数据比较。我们展示了本研究所涉及五种语言的所有共同子集的汇总信息。（BN：广播新闻，CTS：电话会话语音。）

元数据	IIITH	LDC	KGP
时长	15 小时	118 小时	4 小时
说话人	250	378	67
话语	12,663	577	155
语音模式	广播新闻和 CTS	CTS	BN
口音和方言	大多为标准	存在变体	标准
环境	录音室，现实世界	现实世界	录音室
性别比例	均衡	未保持平衡	均衡
音频声道	Mono	立体声	Mono
音频格式	16 kHz (.wav)	8 kHz (.flac)	8 kHz (.wav)

小时语音数据，而LDC规模最大，每种语言约有24小时音频数据。IIITH和KGP的数据大多收集自广播新闻（BN）。LDC数据包含来自NIST LRE 2011 数据集的电话会话语音（CTS）录音。这些语料库中的口音和方言也各不相同。例如，IIITH和KGP语料库中的孟加拉语话语包含的方言主要在印度使用，而LDC语料库包含的方言则在孟加拉国不同地区使用。

IIITH 和 KGP 语料库具有专门的评估协议，包括用于训练集和测试集的数据划分。对于 LDC，数据此前未进行划分。我们随机抽取 80% 的音频录音作为训练数据，其余部分用于测试。LDC话语是立体声音频，每个声道分别对应电话通话中的两位说话人之一。对于每种语言，所有电话通话录音中都有一位共同说话人。为确保说话人不重叠划分，我们在 LDC 测试集中仅考虑排除共同说话人的音频声道。对于所有语料库，训练话语进一步按照 80 : 20 的比例划分为说话人不重叠的训练集和验证集。

3.2. 数据预处理与特征提取

首先，我们对语料库进行处理，使其采样率统一为 8 kHz，音频格式统一为 16位PCM（.wav）。然后，我们使用基于能量的语音活动检测器（VAD） [3]去除非语音区域。最后，我们计算 20 维梅尔频率倒谱系数（MFCCs），使用 20 个滤波器，并对每个话语进行倒谱均值减法（CMS）。提取 MFCC 时，我们使用窗长为 25 ms、帧移长度为 10 ms 的汉明窗。虽然梅尔滤波器组（MFB）特征常用于基于神经网络的分类器，但也有多项研究 [37, 41, 81] 使用 MFCCs 作为输入，应用于我们工作中使用的基于时延神经网络的分类器。值得注意的是，针对所用三个语料库的现有研究 [92, 53, 52, 11] 也主要使用 MFCCs 报告其性能。我们将训练语音话语分割为有效语音时长为 3 s 的片段，以生成训练样本。对于时长超过 3 s 的话语，我们将其分割为多个片段，并丢弃剩余部分。如果话语时长不足 3 s（这种情况极为少见），则丢弃该话语。测试话语也采用类似方式分割为时长为 3 s 的片段。

3.3. 架构描述

基线时延神经网络：作为基线参考，我们采用 x-vector时延神经网络 [82] 架构，并按照我们在 [17] 中的初步研究，使用经过 CMS后处理的 MFCC 特征进行训练。该模型由五个时延神经网络层、一个统计池化（Stat pooling）层和两个全连接（FC）层组成 [81]。表2展示了具体的架构细节。

表2：基线 x-vector时延神经网络的架构详情，其中输入特征维度为 F，时间维度为 T，类别数为 L。

层	上下文	输入维度	输出维度
时延神经网络-ReLU	$[t-2:t+2]$	$F*5\times T$	$512\times T$
时延神经网络-ReLU	$\{t-2、t、t+2\}$	$512*3\times T$	$512\times T$
时延神经网络-ReLU	$\{t-3,t,t+3\}$	$512*3\times T$	$512\times T$
时延神经网络-ReLU	$\{t\}$	$512\times T$	$512\times T$
时延神经网络-ReLU	$\{t\}$	$512\times T$	$1500\times T$
统计池化	T	$1500*T\times 1$	3000×1
全连接-ReLU	-	3000×1	512×1
全连接-ReLU	-	512×1	512×1
全连接-Softmax	-	512×1	$L\times 1$

ECAPA-TDNN：我们采用了 ECAPA-TDNN架构 [16], , 该架构在说话人验证任务中优于其他基础时延神经网络架构 [16]。近期，该架构也被用于语言识别任务 [37, 85]。ECAPA-TDNN 模型以多种方式改进了基础 x 向量架构。模型使用三个基于压缩激励（SE）的残差块（SE-Res2）代替时延神经网络层。利用这三个块的输出进行多层特征聚合（MFA）。随后，将聚合后的输出输入通道与上下文注意力池化层（注意力池化），再接入全连接层。我们对一维卷积（一维卷积）层使用批归一化（广播新闻），并对线性层采用 0.25 的随机失活率。表3详细介绍了该架构。

时延神经网络架构使用 Python 的 PyTorch 库实现 [66] 。我们采用批量大小为 32，并使用 AdamW优化器 [50], 和基于平台期降低学习率的调度器，初始学习率设为 0.001。最大训练轮次设为 30，并在验证损失连续五个训练轮次未改善时应用早停准则。损失函数采用角度间隔（AM）softmax [93] ，缩放参数设为 30，间隔设为 0.2。

表3：ECAPA-TDNN架构的详细信息，其中输入特征维度为 F、时间维度为 T，类别数为 L。（此处，广播新闻表示批归一化。）

层	跳跃连接	输入维度	输出维度
一维卷积+ReLU+广播新闻	-	$F\times T$	$512\times T$
SE-Res2-模块	-	$512\times T$	$512\times T$
SE-Res2-模块	-	$512\times T$	$512\times T$
SE-Res2-模块	-	$512\times T$	$512\times T$
一维卷积+ReLU	层 2、3、4	$512*3\times T$	$1536\times T$
注意力池化+广播新闻	-	$1536*T\times 1$	3072×1
全连接-ReLU	-	3072×1	192×1
全连接-Softmax	-	192×1	$L\times 1$

神经语言识别模型可以通过两种不同方式使用。第一种是端到端架构，将语音作为输入、语言标签作为输出，构建成一个单一模块 [30]。另一种方式是先训练嵌入提取器，然后接入基于概率线性判别分析或余弦相似度的后端分类器 [81]。端到端系统相对简单直接，无须额外训练后端进行评分。另一方面，基于额外后端的方法更加灵活，尤其是在处理集合外语言时 [54]。本研究聚焦于包含固定五种语言的闭集语言识别问题；为简洁起见，我们采用端到端架构。

3.4. 性能评估指标

按照相关文献的做法，我们采用等错误率（EER） [8] 和平均代价（C_{avg}）作为性能评估指标。EER 定义为在虚警率与误拒绝率相等时决策阈值对应的错误率。对于多类别场景，我们计算每种目标语言的一对其余等错误率，并报告其平均值。C_{avg} 是美国国家标准与技术研究院语言识别评测（LRE） [74] 和 OLR 挑战赛 [44]中采用的主要评估指标。其定义如下 [74]:

$$C_{\text{avg}} = \frac{1}{N} \sum_{L_t} \left[p_{\text{Target}} \cdot p_{\text{Miss}}(L_t) + \sum_{L_n} p_{\text{Non-target}} \cdot p_{\text{FA}}(L_t, L_n) \right] \tag{3}$$

其中，L_t 和 L_n 分别表示目标语言和非目标语言。p_{Miss} 和 p_{FA} 分别表示漏检概率和虚警率。p^{Target} 表示目标语言的先验概率，设为 0.5。$p^{Non-target} = （1 – p^{Target}）/（L – 1）$，其中 L 为语言总数。EER 和 C^{avg} 值越低，表示语言识别性能越好。

4. 跨语料库语言识别：领域多样化

低资源语言识别语料库包含的话语数量要少得多。数据采集来源和配置通常都非常有限。因此，在背景、录音设备、室内环境和说话人群体等方面，非语言多样性也受到限制。这些因素会在各数据库中形成语料库特定的非语言偏差，导致跨语料库性能较差 [17]。因此，本节探讨广泛的音频增强方法，以增加低资源语料库的多样性。在语音处理文献中，噪声和非语音添加、音量、音高和速度扰动等音频增强技术，常用于人为增加训练数据的数量和变体 [60, 75, 96]。为解决跨语料库泛化问题，我们重点研究增强方法的广泛应用，原因有二：(i) 增加训练数据有助于避免过拟合问题 [81], ；(ii) 增强训练数据的多样性有助于减少语料库失配。为在训练数据中融入多种随机性来源，我们探讨了若干现有和新颖的音频增强技术。此外，我们提出一种名为最大多样性感知级联增强方法的新概念，并对增强折叠因子进行优化。

4.1. 所探索的增强方法说明

根据音频数据扰动的方式，我们将所研究的增强技术分为八类。我们主要关注直接作用于原始波形的信号级增强类别

表 4：语言识别框架中应用的不同信号级和特征级增强类别及其各自子类别的汇总。

加性非语音 增强 (A^1)	信号参数 扰动 (A^2)	带窝 增强 (A^3)	环境 变体 (A^4)	语音增强 (A^5)	音频编码 (A^6)	从.....恢复 有损编解码器 (A^7)	特征级 增强 (A^8)
喧闹声 音乐 噪声	音高 偏移 速度 音量	下限截止频率 上限截止频率 电话 信道 BPF	智能手机 现场大厅 广播 黑胶唱片录音	谱减法 基于最小均方误差 去混响	A律 IMA自适应差分脉冲编码调制 OKI自适应差分脉冲编码调制 μ律	.aac .gsm .mp3 .ogg Opus音频格式 .wma	SpecAug Mixup

。此外，我们也研究了特征级增强。每个类别下还有不同的子类别。不同的增强类别是根据我们初步调查中的语料库分析结果选定的 [17]，它们引入的多样性有助于减少语料库失配。我们首先研究了各类别对语言识别性能的影响。随后，我们汇集并采样了不同类别的话语，以积累它们所具备的不同领域多样化方式。表 4 总结了所采用的增强类别。以下将详细介绍这些增强类别。

4.1.1. 加性非语音增强

我们的初步研究中对长期平均频谱（LTAS）和信噪比（SNR）直方图的分析表明，不同低资源语言识别语料库之间的信道效应和背景可能有所不同 [17]。因此，我们在训练语句中加入加性噪声、babble噪声和音乐片段等非语音效应。此类增强方法沿用自 [82]，并通过 Kaldi工具的 [68] VoxCeleb 配方 [1] 实现。该方法使用来自 MUSAN语料库 [80] 的非语音样本。

4.1.2. 信号参数扰动

不同语料库中的话语也可能因说话风格而异。例如，不同说话人的音高、语速和响度可能各不相同。为减少这些方面的语料库失配，我们使用音频退化工具箱 [55] 随机改变不同的信号参数。

- 音高：随机偏移 $[-4, 4]$ 个半音，同时保持话语长度不变。
- 移位：将训练语句随机分成两部分并交换。
- 速度：随机选择 $\Gamma \in [-15, +15]$ 来改变播放速度。增强音频的长度 (l) 变为 $l' = ((100 - \Gamma)/100)l$。$\Gamma > 0$ 会提高播放速度，而 $\Gamma < 0$ 会降低播放速度 [55]。
- 音量：在 $[-30, 40]$ dB 范围内随机调整增益。

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

4.1.3. 带宽增强

根据 [17], 中的长期平均频谱（LTAS）比较，我们发现不同语料库的频谱在不同频率范围内可能存在显著差异。因此，我们使用截止频率随机变化的滤波器处理话语。

- 上限截止频率增强：使用固定的下截止频率（f_L）20 Hz 进行滤波。上截止频率（f_U）从 $f_U \in [2500, 3500]$ Hz 中随机选取。
- 下截止频率增强：使用固定的 $f_U = f_s/2$（f_s 表示采样频率）进行滤波，并随机选择 $f_L \in [50, 200]$ Hz。
- 电话信道带通滤波（BPF）：为了复现旧式模拟电话信道中出现的滤波效果（遵循表 1 中的 LDC 元数据），我们使用带通滤波器对音频进行滤波，其中 $f_L = 300$ Hz 固定，$f_U \in [3000, 4000]$ Hz 可变。

4.1.4. 卷积不同的房间环境响应

根据表 1，IIITH 和 KGP 的话语大多采集自录音室。LDC 数据则是在户外和较小的房间内录制的。为了模拟不同的录音环境，我们使用 Kaldi 工具 [68]，将话语与从 RIR 语料库中随机选取的房间脉冲响应进行卷积。此外，我们使用音频退化工具箱 [55],, 将话语环境转换为黑胶唱片录音、现场大厅录音、智能手机播放和无线电广播。

4.1.5. 语音增强

[17] 中的信噪比直方图分析表明，不同语料库中的话语可能受到不同程度的噪声干扰。为减少因噪声水平变化而导致的不匹配，我们提出将语音增强作为一种增强类别。使用 Voicebox工具 [7] 随机对话语应用三种语音增强算法中的一种：谱减法 [4], 、基于对数MMSE噪声估计的增强 [27], 和去混响 [20]。

4.1.6. 更改音频编码器

不同语料库中的中间数据录制、处理和存储可能采用不同的编码格式。继 [25], 中的增强之后，我们探索了几种用于数据增强的编码格式。我们随机将话语的编码格式转换为线性脉冲编码调制（L-PCM）、A律、μ律、IMA 和 Oki自适应差分脉冲编码调制（ADPCM）中的一种。[2]

4.1.7. 从有损音频编解码器中恢复

有损编解码器可在一定范围内降低音频保真度，以实现数据压缩。在中间处理过程中，各语料库中的话语可能以不同的有损编解码器格式存储，之后再转换为无损的 .wav（用于 IIITH 和 KGP）或 .flac（用于 LDC）。通过有损编解码器进行增强，可利用依赖编解码器的部分信息损失，使话语更加多样化。使用 SoX 库，[3] 我们随机将话语转换为不同的有损编解码器格式，例如 .aac、.mp3、.gsm 和 .opus，然后再将其还原为 .wav。

[2] https://github.com/FFmpeg/FFmpeg
[3] https://github.com/rabitt/pysox

4.1.8. 特征级增强

迄今讨论的增强方法均应用于信号级。我们还探索了两种广泛使用的增强方法在特征级应用时的适用性。

SpecAug [65]：这是一种特征级增强方法，通过以下方式随机修改滤波器组特征：(i) 沿时间轴扭曲；(ii) 时间掩蔽；(iii) 在随机时频位置进行频率掩蔽。该方法增强了模型对时间轴扭曲以及特定频率通道和时间帧中部分信息丢失的鲁棒性 [65]。SpecAug 主要用于频谱图或 MFB。由于采用了掩蔽，标准 SpecAug 会将部分 MFB 值置零，其余值保持不变。然而，在本研究中，我们将 SpecAug 用于 MFCC。具体做法是在 MFCC 提取过程中，直接对作为中间阶段生成的 MFB 应用 SpecAug。在最终的 DCT 阶段进行去相关时，SpecAug 会扰动整个特征矩阵。本研究将 SpecAug 用于 MFCC，主要是为了与基线以及使用 MFCC 的三个语料库所对应的现有 LID 文献进行公平比较。尽管如此，此前也有若干研究将 SpecAug 应用于 MFCC [21, 72, 87, 97]。

Mixup [99]：该增强方法通过对随机选取的训练样本（X_i，y_i）和（X_j，y_j）进行线性插值，生成新的训练样本（$\mathrm{X}_{\mathrm{mix}}$，$\mathrm{y}_{\mathrm{mix}}$）。其定义如下：

$$\begin{aligned} \mathbf{X}_{\mathrm{mix}} &= \theta \mathbf{X}_i + (1-\theta)\,\mathbf{X}_j, \\ y_{\mathrm{mix}} &= \theta y_i + (1-\theta)\,y_j, \end{aligned} \tag{4}$$

其中，X_i和 X_j表示两个独立样本的特征输入，y_i和 y_j是以独热编码表示的相应标签。 θ 是混合参数，服从贝塔分布，其两个形状参数相等且介于 0 和 ∞ 之间。Mixup 会生成软标签 $\mathrm{y}_{\mathrm{mix}}$，从而增强对错误标签和对抗样本 [99]的鲁棒性。

4.2. 优化增强比例

数据增强方法有助于提升时延神经网络（TDNN） [82]等神经网络架构的泛化能力。研究表明，采用多种方法进行数据增强有助于训练更大的模型 [25]。然而，我们推测，为了解决低资源语料库训练数据不足的问题，如果使用数量过多的增强样本，可能会增加计算量。此外，与原始样本相比，增强语句过多时，语言识别模型可能开始学习声学变化，而不再学习原始训练数据中存在的真正语言判别线索 [83]。尽管在探索不同音频增强技术方面已取得显著进展，但我们发现，现有文献 [25, 36, 96] 并未对训练所需增强的音频数据量进行全面分析。在本研究中，我们深入探索增强语句与组合（增强 + 原始）训练集中的原始语句之间的最佳比例（以同语料库和跨语料库语言识别性能为依据）。

假设对于任意第 i 个语料库，未增强原始训练话语的总数为 N_O^i。对于每个信号级增强类别 A^k（其中 $\mathrm{k} \in \{1, 2, \cdots, \mathrm{K}\}$），我们为其所有 S^k 个子类别各创建一个增强实例。因此，对应于 A^k 的增强语句总数为 $\mathrm{M}^k = \mathrm{S}^k \mathrm{N}_O^i$。如果从这 M^k 个话语中随机抽取 η^k 比例，那么增强语句总数 (N_A^i) 为 $\mathrm{N}_A^i = \mathrm{P}_{k=1}^K \mathrm{M}^k \eta^k$。

如果 $\eta^k = \gamma/(KS^k)$（其中 γ 为非零正实数），则合并后的增强训练集中的话语总数 (N^i_{Tot}) 为:

$$N^i_{\text{Tot}} = N^i_O + N^i_A = (1+\gamma)N^i_O \tag{5}$$

在本研究中，γ 称为增强折叠因子。它表示组合训练集中增强语句与原始未增强训练语句的比率。我们针对 $\gamma = 0.5$、1、2、3 和 4 创建不同的组合训练集。除了评估各增强类别的语言识别性能外，我们还针对不同的增强折叠因子训练独立的语言识别模型，并分析其性能。

4.3. 级联增强类别

我们采用的领域多样化方法旨在重现未见跨语料库评估语句中存在的非语言多样性，以提升泛化能力。为此，我们创建更逼真的增强音频，以模拟现实世界中不同的声学场景。因此，除了传统上分别探索不同音频增强类别的方法外，我们还提出了一种名为最大多样性感知级联增强的算法：该算法将多种数据增强类别结合使用，并根据引入的多样性选择类别。

文献中有几种增强方法会通过多个信号处理阶段对数据进行扰动。例如，基于房间环境的增强（A^5）包括混响效果，随后进行噪声添加。然而，这类示例十分有限，其潜力也鲜少得到探索。在本研究中，我们提出了最早一批理论与实验研究之一，探讨如何高效地级联多个增强类别。在级联增强中，类别 A^i 和 A^j 的信号处理步骤可以级联，形成新的增强类别 $A^{\{i,j\}}$。鉴于所研究的增强类别涉及多种信号处理操作，我们假设级联顺序是非交换的，即 $A^{\{i,j\}}$，$A^{\{j,i\}}$。我们可以将两个或更多不同的增强类别串联起来，生成多样化的增强音频数据。在此过程中，我们需要权衡话语多样化的程度与保留增强话语的语义语言标签之间的关系。我们通过将级联限制为仅包含两个信号级增强类别来应对这一权衡。

我们的方法依赖于不同增强类别之间语句嵌入的距离。我们训练语言识别模型，并提取属于信号级增强类别 $A^k \; \forall \, k \in [1,K]$的训练话语嵌入 Z^k。我们使用最终 softmax层之前线性层的输出，提取 192 维的话语级嵌入（Z^k）。随后，我们计算嵌入分布之间的多样性 $D^{\{i,j\}} = D||p(Z^i), p(Z^j)||$。其中，$i$、$j \in [1, K]$ 表示两个不同的信号级增强类别，且 i , j。

为衡量多样性，我们采用三种指标 [106]：余弦距离、KL散度和对称KL散度。对于这三种距离度量，我们计算原始语句（记为 A^0）与来自 A^k 的增强语句之间的嵌入距离（$D^{\{0,k\}}$）。随后，我们对 $D^{\{0,k\}}$ 值与相应的 c^k_{avg} 分数进行相关性分析。这里，c^k_{avg} 表示同语料库语言识别性能（在验证数据上），其中训练数据由类别 A^k 的增强语句与原始训练语句组合而成。根据实验结果（见第6节 6.5），我们选择 KL散度作为多样性度量指标，因为它呈现出最强的负相关。负相关

算法1 最大多样性感知级联增强。

过程（将 A^i 和 A^j 级联为 $A^{\{i,j\}}$）　▷ i、j ∈ [1,K] 且 i、j

　对于 i = 1 : K, 执行

　　$Z^i \leftarrow A^i$　▷ Z：时延神经网络嵌入

　　对于 m = 1 : K, 执行

　　如果 i , m, 则

　　$D^{\{i,m\}}\,KL \leftarrow D_{KL}(p(Z^i), p(Z^m))$　▷ D_{KL} : KL散度

　　如果 $D^{\{i,j\}}{}_{KL} > D^{\{i,m\}}{}_{KL}$, $\forall$ m ∈ [1,K] ≠ j, 则

　　级联 A^i 后接 A^j。

意味着领域多样性越高，性能越好（即 C_{avg} 越低）。此外，与另外两种距离度量不同，KL散度不具对称性，$D^{\{i,j\}}KL$，$D^{\{j,i\}}KL$。这也与级联增强方法的非交换性相一致。

对于 K 个增强类别，总共可能有 $({}^{K}P_2)$ 个级联配对。我们避免重复相同的增强类别。因此，有效的排列数为 $({}^{K}P_2) - K$。相应地，我们有一个成对KL散度得分 $D^{\{i,j\}}{}_{KL}$，其中 i、j ∈ [1, K]，且 i , j。两个不同增强类别的嵌入之间的KL散度得分，表明它们为训练数据带来的多样性变化。为了最大限度地提升级联增强的多样性，对于每个增强类别 i ∈ [1, K], , 我们将其与增强类别 j ∈ [1, K] ≠ i 级联，使得：$D^{\{i,j\}}\,KL > D^{\{i,m\}}\,KL$, ∀ m ∈ [1, K] ≠ j。生成的级联增强话语记为 $A^{\{i,j\}}$。算法1概述了我们提出的、用于生成多样性感知程度最高的级联增强话语的方法。

我们还将两种特征级增强技术 SpecAug 和 Mixup 分别级联在不同的信号级增强方法之上。这样做是因为信号级增强和特征级增强所涵盖的音频扰动来源不同。因此，将二者级联可以进一步提升训练数据的多样性。

5. 跨语料库语言识别：域泛化

非语言因素上的跨语料库失配可被视为一个域偏移问题 [10]。由协方差偏移和概念偏移引起的域偏移会改变训练（源）数据样本与测试（目标）数据样本之间的特征—标签联合分布 $p(\mathcal{X}^{src}\mathcal{Y}^{src})$、$p(\mathcal{X}^{tar}\mathcal{Y}^{tar})$ [58]。本节将从域泛化 [34] 的角度，处理跨语料库域偏移问题。

5.1. 域自适应与域泛化

机器学习文献中一种常见的方法是在训练期间纳入目标域样本中的信息，以缓解领域偏移问题。在源域上训练的模型通常会针对特定的目标域进行适配 [76]。这种方法称为领域自适应 [100, 95]。在我们的语境中，可以将训练语料库视为源域，并将其余两个语料库视为各自独立的目标域。然而，鉴于第2节所述跨语料库评估的限制，我们不使用目标域的任何数据。此外，针对某一特定目标域进行适配，并不能保证模型能够泛化到现实世界部署期间的任何未见测试域 [19]。考虑到这些因素，我们选择了另一种利用

源域信息并提升语言识别模型泛化能力的方法。这种方法称为域泛化 [103]。域泛化的概念最早是在计算机视觉领域正式提出的 [6]。对于领域偏移问题 [10]，当我们不针对任何特定目标域，而是希望在现实世界未见条件下实现鲁棒性时，域泛化正逐渐成为研究界关注的一种现实解决方案。

5.2. 伪域：解决多个源域问题

在语音处理文献中，由于收集多源标注数据存在困难，域泛化方法的研究相对较少 [28]。因此，我们首先提出一种将多个源域的数据引入框架的解决方案。在第4.1节中，我们讨论了用于领域多样化的不同信号级增强类别。在此，我们提出利用这些增强类别以及原始条件作为多个源域，从域泛化的角度解决跨语料库语言识别问题。我们将这些合成域称为伪域，因为它们是从源域数据中人工创建的。对于当前研究而言，由于低资源语言的训练数据有限，在域泛化中利用音频增强更加合适。

将每个伪域表示为 $\mathcal{D}^{\text{dom}} = \{(\mathrm{X}_i^{\text{dom}}, \mathrm{y}_i^{\text{dom}})\}^{N_{\text{dom}}}_{i=1}$，其中 $\text{dom} \in [0, \mathrm{K}]$，K 表示信号级增强类别的总数，$\mathrm{N}_{\text{dom}}$ 表示相应伪域的训练样本数量。假设话语 $\{(\mathrm{X}_i^{\text{dom}}, \mathrm{y}_i^{\text{dom}})\}$来自域特定联合分布 $\mathrm{p}(\mathcal{X}^{\text{dom}}, \mathcal{Y}^{\text{dom}})$。在不适应任何特定目标域数据 $\mathcal{D}^{\text{tar}} = \{(\mathrm{X}_i^{\text{tar}}, \mathrm{y}_i^{\text{tar}})\}^{N_{\text{tar}}}_{i=1}$ 的情况下（其中 N_{tar} 表示目标域的样本数量），域泛化的目标是泛化到未见过的目标域分布 $\mathrm{p}(\mathcal{X}^{\text{tar}}, \mathcal{Y}^{\text{tar}})$，其中 $\text{tar} \in [\mathrm{K}+1, \infty)$ [34]。在以下小节中，我们讨论了用于解决跨语料库语言识别问题的不同域泛化方法。

5.3. 域对抗式ECAPA-TDNN架构

本文进一步扩展了当前最先进的 ECAPA-TDNN架构，加入了一个额外的对抗式域分类器分支，用于学习域不变语言信息。如图 2 所示，所提出的架构包含一个编码器（$\mathcal{E}$），该编码器由帧级 SE-Res2模块和注意力池化层组成。对于每个话语，编码器将输入特征向量 X 映射为 Z。在编码器之后，我们采用了两个片段级分类分支：语言识别分支（$\mathcal{D}_{\text{lang}}$）和域分支（$\mathcal{D}_{\text{dom}}$）。$\mathcal{D}_{\text{lang}}$ 由两个全连接层组成。对于第 i 个话语，它输出语言后验概率 $\mathrm{p}(\mathrm{y}|\mathrm{X}_i; \theta_E, \theta_L)$。其中，$\theta_E$ 和 θ_L 分别表示编码器和语言识别分支的可学习参数。额外的域分类器分支 $\mathcal{D}_{\text{dom}}$（其可学习参数记为 θ_D）将伪域后验概率 $\mathrm{p}(\mathrm{y}^{\text{dom}}|\mathrm{X}_i; \theta_E, \theta_D)$预测为一个 $\mathrm{K}+1$ 类 Softmax 向量，其中包含 K 个信号级增强类别以及原始未增强话语这一额外类别。这里，我们不在域分类器中纳入子类别，因为它们都会以相似的方式扰动音频。此外，由于子类别导致相似输出类别过多，可能会使域分类任务变得困难，进而阻碍主要的语言识别任务。这里，$\mathcal{Y}^{\text{dom}} \in \{0, 1, ..., \mathrm{K}\}$ 表示伪域标签空间。$\mathcal{D}_{\text{dom}}$ 分支由一个一维卷积层和两个全连接层组成。它通过梯度反转层（GradRev）[24] 与 $\mathcal{E}$ 模块相连，该层会在反向传播过程中翻转流入编码器的梯度符号。训练期间，将伪域数据 $\{\mathrm{X}^{\text{dom}}, \mathrm{y}, \mathrm{y}^{\text{dom}}\}\ \forall\ \text{dom} \in [0, \mathrm{K}]$ 作为输入。我们通过优化定义如下的组合对抗损失函数（$\mathcal{L}_{\text{adv}}$）来训练模型：

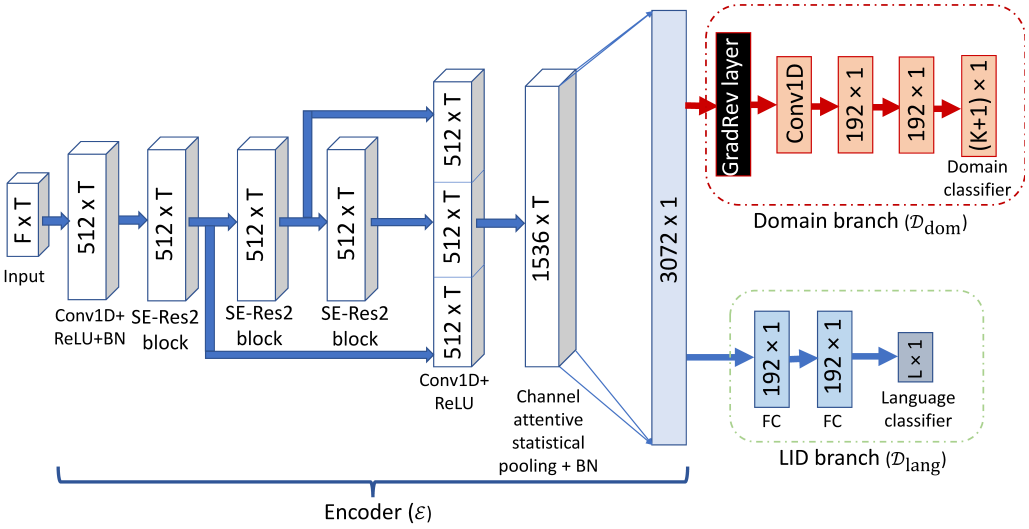

图2. 用于域泛化的对抗式 ECAPA-TDNN 架构。这里，K = 7（其中“+1”表示原始未增强类别）表示对应于每个信号级类别的伪域总数。L 表示目标语言的数量；在本闭集语言识别研究中，目标语言数为五种。

$$\begin{aligned}\mathcal{L}_{adv}(\theta_E,\theta_L,\theta_D) = \sum_i[(1-\lambda_{\text{adv}})\mathcal{L}_{\text{lang}}(y_i,\mathcal{D}_{\text{lang}}(\mathcal{E}(\mathbf{X}_i));\theta_E,\theta_L)\\ -\lambda_{\text{adv}}\ \mathcal{L}_{\text{dom}}(\mathcal{D}_{\text{dom}}(y_i^{\text{dom}},\mathcal{E}(\mathbf{X}_i));\theta_E,\theta_D)].\end{aligned} \tag{6}$$

这里，i 表示各个话语的索引，$\mathbf{X}_i \in \mathcal{X}$。总损失（$\mathcal{L}_{\text{adv}}$）由语言分类损失 $\mathcal{L}_{\text{lang}}$ 和域分类损失 $\mathcal{L}_{\text{dom}}$组成。参数$\lambda_{\text{adv}} \in [0,1]$ 决定总损失中两种损失的权重。通过优化极大极小问题，$\mathcal{L}_{\text{lang}}$ 训练 θ_L 和 θ_E，以提高语言识别准确率。同时，θ_D 经过估计以提升伪域识别性能。梯度反转层会翻转 $\delta\mathcal{L}_{\text{dom}}/\delta\theta_D$的符号，使得 θ_E 能够学习域不变的语言判别信息。

5.4. 用于跨语料库泛化的多任务学习

我们在跨语料库语言识别框架中探索了多任务学习（MTL）[101, 73, 9]，旨在学习能够提升泛化能力的伪域感知表征。联合学习多种特征或多项任务已被证明能够提升跨信道语言识别测试样本的鲁棒性 [43]。借助 MTL，我们的语言识别模型还联合学习一项八分类域分类任务 [103]。因此，语言识别模型有望捕获能够在多个伪域间泛化的域特定语言判别信息。在此实验中，对 ECAPA-TDNN 架构的扩展与对抗式方案（见第5.3节）类似，区别在于移除了梯度反转层 [1]。待优化的总损失可表示为

$$\begin{aligned}\mathcal{L}_{\text{MTL}}(\theta_E,\theta_L,\theta_D) = \sum_i[\mathcal{L}_{\text{lang}}(y_i,\mathcal{D}_{\text{lang}}(\mathcal{E}(\mathbf{X}_i));\theta_E,\theta_L)\\ +\lambda_{\text{MTL}}\ \mathcal{L}_{\text{dom}}(y_i^{\text{dom}},\mathcal{D}_{\text{dom}}(\mathcal{E}(\mathbf{X}_i));\theta_E,\theta_D)].\end{aligned} \tag{7}$$

在公式 6 和公式 7 中，主要区别在于 $\mathcal{L}_{\text{dom}}$ 分量的符号。因此，基本上，

这两项实验试图采用两种相反的策略来解决域失配问题：域不变和域感知 [1]。在域泛化中，我们无法获得任何特定目标域的信息。因此，这两种相反但互补的域泛化方法，都适用于现实世界应用中可能遇到的不同未见域。

5.5. 基于最大均值差异的域泛化方法

机器学习中最基本的假设是，训练数据和评估数据独立同分布（i.i.d.）。然而，由于领域失配，这一假设往往会被违反。为了提升泛化能力，我们利用伪域，提出的思路是训练语言识别模型，以减少训练条件之间的分布失配。在本研究中，我们探索了最大均值差异（MMD） [32] 域泛化技术。我们研究 MMD 主要出于两个原因。首先，它能有效缩小源数据与目标数据之间的分布偏移 [32]。其次，语言识别领域的研究已成功运用 MMD 来减少领域失配 [22]。

MMD 是用于衡量两个分布之间距离的度量 [32]。在域自适应文献中，通过优化基于最大均值差异的损失（$\mathcal{L}_{\text{MMD}}$），可以最小化源域与目标域之间统计矩的差异 [105]。在本研究中，我们首先将伪域用作目标域，从而在域泛化配置下应用基于最大均值差异的域自适应技术。随后，我们通过提出多域扩展，扩展基于最大均值差异的域泛化模型。在传统的基于最大均值差异的域泛化实验中，我们将未增强的训练语句视为源域（$\mathcal{S}^{\text{src}}$），并将所有伪域（$\mathcal{S}^{\text{dom}}$）视为目标域。

设源域和目标域的特征空间分布分别表示为 $\mathrm{p}^{\text{src}} = \mathrm{p}(\mathcal{X}^{\text{src}})$ 和 $\mathrm{p}^{\text{dom}} = \mathrm{p}(\mathcal{X}^{\text{dom}})$。我们首先计算最大均值差异损失，即映射到再生核希尔伯特空间（RKHS）中的源表示与目标表示的均值之间的平方差 [32] ，其定义如下：

$$d_{\mathcal{H}} = \left\| \sum_{\mathbf{X}_i^{\text{src}} \in \mathcal{S}^{\text{src}}} \phi(\mathbf{X}_i^{\text{src}}) - \sum_{\mathbf{X}_i^{\text{dom}} \in \mathcal{S}^{\text{dom}}} \phi(\mathbf{X}_i^{\text{dom}}) \right\|_{\mathcal{H}}^2 . \tag{8}$$

在公式 8 中，i 表示话语索引，$\mathrm{X}^{\text{src}} \in \mathcal{X}^{\text{src}}$、$\mathrm{X}^{\text{dom}} \in \mathcal{X}^{\text{dom}}$。 $\mathcal{H}$ 表示与半正定核 $\mathcal{K}$对应的再生核希尔伯特空间。最大均值差异损失与语言识别损失联合优化，以增强语言识别模型对域偏移的鲁棒性。实际应用中，假设存在正定核$\mathcal{K}(\mathrm{X}^{\text{src}}, \mathrm{X}^{\text{dom}}) = \langle \phi(\mathrm{X}^{\text{src}}), \phi(\mathrm{X}^{\text{dom}}) \rangle$，通过公式 9 计算最大均值差异损失的经验估计值。这里， ϕ 将输入从特征空间映射到再生核希尔伯特空间，〈、〉算子表示内积。

$$\hat{d}_{\mathcal{H}} = \left\| \frac{1}{N_{\text{src}}} \sum_{i=1}^{N_{\text{src}}} \phi(\mathbf{X}_i^{\text{src}}) - \frac{1}{N_{\text{dom}}} \sum_{i=1}^{N_{\text{dom}}} \phi(\mathbf{X}_i^{\text{dom}}) \right\|_{\mathcal{H}}^2 \tag{9}$$

这里，N_{src} 和 N_{src} 分别表示源域和目标域话语的数量。进一步简化公式 9 可得：

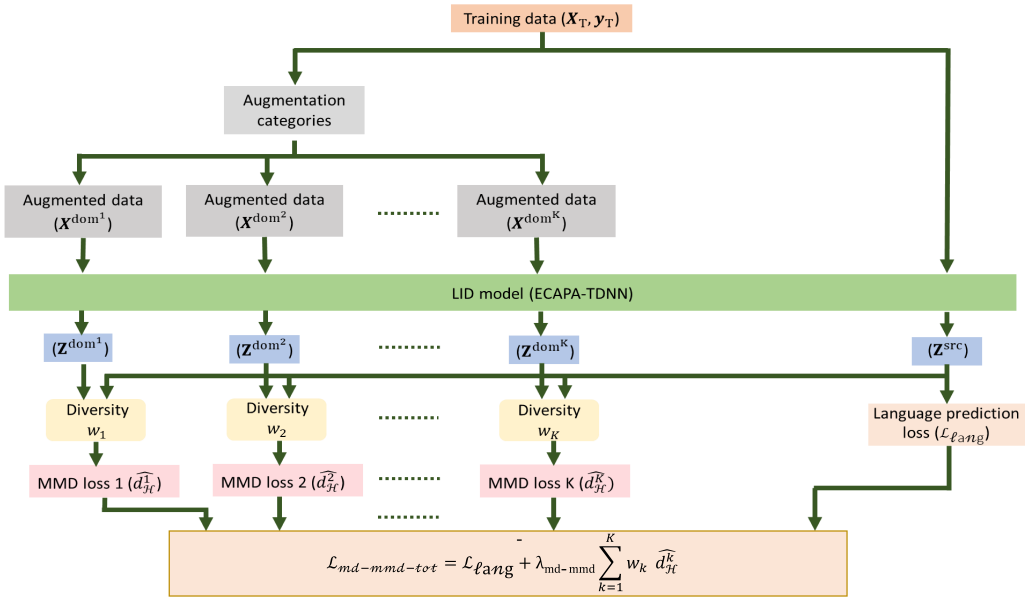

图3. 基于 MD-MMD 的 ECAPA-TDNN 模型的工作原理，其中 $\mathcal{L}_{\text{lang}}$ 表示 AM-softmax 语言预测损失，$\hat{\mathrm{d}}^k_{\mathcal{H}}$ 表示各个多域最大均值差异损失分量。

$$\hat{d}_{\mathcal{H}} = \frac{1}{N_{\text{src}}^2}\sum_{i=1}^{N_{\text{src}}}\sum_{j=1}^{N_{\text{src}}}\phi(\mathbf{X}_i^{\text{src}})\phi(\mathbf{X}_j^{\text{src}}) + \frac{1}{N_{\text{dom}}^2}\sum_{i=1}^{N_{\text{dom}}}\sum_{j=1}^{N_{\text{dom}}}\phi(\mathbf{X}_i^{\text{dom}})\phi(\mathbf{X}_j^{\text{dom}}) - \frac{2}{N_{\text{src}}N_{\text{dom}}}\sum_{i=1}^{N_{\text{src}}}\sum_{j=1}^{N_{\text{dom}}}\phi(\mathbf{X}_i^{\text{src}})\phi(\mathbf{X}_j^{\text{dom}}) \tag{10}$$

在最大均值差异损失中使用多核扩展可提升分类性能 [48]。在本研究中，我们使用核函数 $\mathcal{K}$，它由 M 个具有不同方差的径向基函数（RBF）组合而成：

$$\mathcal{K}(\mathbf{X}_i, \mathbf{X}_j) = \sum_{m=1}^{M} exp\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^T(\mathbf{X}_i - \mathbf{X}_j)}{\sigma_m^2}\right) \tag{11}$$

其中，

$$\sigma_m^2 = \frac{2^m}{N_{\text{src}}N_{\text{dom}}}\sum_{i=1}^{N_{\text{src}}}\sum_{j=1}^{N_{\text{dom}}}(\mathbf{X}_i - \mathbf{X}_j)^{\top}(\mathbf{X}_i - \mathbf{X}_j) \tag{12}$$

组合损失定义为

$$\mathcal{L}_{\text{mmd-tot}} = \mathcal{L}_{\text{lang}} + \lambda_{\text{mmd}}\,\hat{d}_{\mathcal{H}}. \tag{13}$$

在公式 13 中，λ_{mmd} 是用于确定语言识别损失和 MMD损失相对权重的标量。联合优化在学习语言判别表征的同时，使源域和目标域的表示更加相似。

在式 13 中，所有伪域都被统一视为单个目标域 $\mathcal{S}^{\text{dom}^\wedge}$，并对应单个 MMD 损失分量 $\mathrm{d}_{\mathcal{H}}$。为了更高效地利用伪域之间更细粒度的域差异，我们提出将传统最大均值差异方法扩展为多域最大均值差异（MD-MMD）损失。所提出的 MMD 方法能够高效地纳入

每个伪域各自不同的多样性来源。对于每个伪域 $\mathcal{S}^{\text{dom}^k}$，其中 $\text{k} = [1 : \text{K}],$，我们分别计算 MMD 损失分量 $\hat{\text{d}}_{\mathcal{H}}^{k}$:

$$\hat{d}_{\mathcal{H}}^{k} = \left\| \frac{1}{N_{\text{src}}} \sum_{i=1}^{N_{\text{src}}} \phi(\mathbf{X}_i^{\text{src}}) - \frac{1}{N_{\text{dom}}^{k}} \sum_{i=1}^{N_{\text{dom}}^{k}} \phi(\mathbf{X}_i^{\text{dom}^k}) \right\|_{\mathcal{H}}^{2} \tag{14}$$

此处，N_{dom^k} 表示 $\mathcal{S}^{\text{dom}^k}$的训练话语数量。随后使用 $\mathcal{L}_{\text{lang}}$同时优化每个组成部分。MD-MMD 的总损失函数为:

$$\mathcal{L}_{\text{md-mmd-tot}} = \mathcal{L}_{\text{lang}} + \lambda_{\text{md-mmd}} \sum_{k=1}^{K} w_k \, \hat{d}_{\mathcal{H}}^{k} \tag{15}$$

在公式 15 中，w_k 是 MD-MMD 损失中各个伪域的相对权重。一种简单的方法是将 $\text{w}_k = 1/\text{K} \;\; \forall \, \text{k} \in [1,\text{K}]$。然而，不同伪域在源域中提供的域多样性程度各不相同。因此，可为与源域相比多样性度量最高的域赋予最大权重。在所提出的方法中，w_k 根据源域中相应的归一化域多样性度量（见第 5.3 节）设定:

$$w_k = \left[\mathbf{D}_{KL}^{\{0,k\}} / \sum_{k=1}^{K} \mathbf{D}_{KL}^{\{0,k\}} \right]. \tag{16}$$

所提出的 MD-MMD 模型的详细工作原理如图3所示。

6. 结果与分析

本节将呈现并分析本研究中不同实验的语言识别性能结果。按照 NIST LRE [74] 和 OLR 挑战赛 [44], 的做法，语言识别性能以等错误率 (%) / $\text{C}_{\text{avg}} * 100$ 表示（表中记为 EER/C_{avg}）。评估时，我们主要关注跨语料库与同语料库性能指标之间的不匹配。对于所提出的方法，我们还研究了与基线相比，同语料库语言识别性能的相对提升。首先，我们比较 x-vector基线时延神经网络和 ECAPA-TDNN架构的语言识别性能。然后，分别评估各增强类别的语言识别性能。接着，训练并评估前文讨论的不同领域多样化方法。最后，将多样化数据用作伪域，并评估域泛化扩展的语言识别性能。

6.1. 基线语言识别性能

我们首先使用基线 x-vector时延神经网络和 ECAPA-TDNN 架构训练语言识别模型。相应的同语料库和跨语料库评估结果见表 5。结果表明，在大多数评估情况下，ECAPA-TDNN 的表现优于 x-vector时延神经网络。我们发现，仅在使用 KGP测试数据进行评估的部分情况下，x-vector 的性能略有提升。与 x-vector 相比，ECAPA-TDNN 中建模的额外改进可能会降低其对 KGP测试数据中非语言不匹配的鲁棒性。KGP测试子集包含

表 5：时延神经网络架构的语言识别性能评估，等错误率（%）/ $C_{avg} * 100$。

训练语料库	基线时延神经网络			ECAPA-TDNN		
	IIITH测试集	LDC测试集	KGP测试集	IIITH测试集	LDC测试集	KGP测试集
	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}
IIITH	12.40 / 13.82	43.10 / 46.23	30.30 / 29.91	9.74 / 11.51	42.43 / 44.82	34.83 / 32.62
LDC	50.08 / 47.34	22.95 / 26.40	45.60 / 45.92	46.35 / 43.21	21.86 / 25.70	42.95 / 39.59
KGP	37.55 / 34.91	51.80 / 46.72	11.18 / 10.82	36.52 / 33.59	47.25 / 45.49	12.36 / 8.75

表6：IIITH 训练数据中各单项信号级增强类别对应的语言识别性能（等错误率(%) / $C_{avg} * 100$）。

测试语料库	不使用增强	非语音添加（A^1）	信号参数扰动（A^2）	带宽增强（A^3）	环境变体（A^4）	语音增强（A^5）	音频编码（A^6）	有损编解码器恢复（A^7）
IIITH	9.74 / 11.51	8.43 / 8.73	10.12 / 11.76	9.76 / 11.56	9.48 / 9.46	9.21 / 9.06	8.93 / 9.01	9.63 / 11.42
LDC	42.43 / 44.82	40.33 / 36.78	41.15 / 43.64	43.77 / 46.11	39.62 / 37.80	39.84 / 38.32	39.32 / 36.44	42.71 / 44.70
KGP	34.83 / 32.62	33.23 / 39.01	39.87 / 38.70	42.06 / 37.32	37.08 / 33.72	35.06 / 31.91	32.36 / 29.98	38.26 / 35.59

每种语言仅有 60 个语块（每个语块时长为 3 秒）。相比之下，IIITH测试集共包含 4878 个时长为 3 秒的语块，而 LDC测试集共包含 24617 个时长为 3 秒的语块。由于 KGP测试集规模极小，可能存在样本偏差，导致 ECAPA-TDNN 的性能较差。尽管如此，由于 ECAPA-TDNN 在大多数语言识别评估中的性能更佳，我们在后续实验中仅考虑基于 ECAPA-TDNN 的语言识别模型。对于这两种架构，我们都观察到同语料库评估与跨语料库评估之间存在显著的性能差异。跨语料库语言识别实验的等错误率超过 30%，表明语言预测不可靠。较高的等错误率也反映出，要在不同语言识别语料库之间实现良好的泛化是一项具有挑战性的任务。

6.2. 语言识别性能：各增强类别

对于七种信号级增强类别 A^k，我们针对每个相应的子类别，创建训练语句的一个增强副本。随后，对于每种增强类别，我们随机抽取增强语句（扩增倍数为 $\gamma = 1$），并将其与原始未增强训练数据（A^0）合并。因此，我们创建了七组增强音频数据，并分别训练语言识别模型。表6展示了各信号级增强类别的语言识别性能对比。为便于比较，表6还列出了使用未增强的原始训练数据训练的 ECAPA-TDNN 模型的语言识别性能。

对于这三个训练语料库，我们发现所探索的增强类别带来的语言识别性能提升幅度相近。因此，在表6以及后续的域多样化实验中，我们仅展示了使用 IIITH 训练的情形下的语言识别性能。从表6可以看出，在同语料库评估中，非语音添加类别的表现最佳。对于 LDC 和 KGP 的跨语料库评估，编解码器编码增强带来的性能提升最大。所提出的基于语音增强的增强方法和环境增强，在两项跨语料库评估中也持续展现出良好的性能提升。随后，我们尝试将这些跨语料库评估结果与现有元信息联系起来。与 IIITH 和 KGP 的广播新闻演播室录音不同，LDC 数据包含在室内和室外环境中录制的电话交谈。因此，多样化的背景噪声

表7：使用 IIITH 作为训练语料库时，数据增强折叠因子对语言识别性能的影响（等错误率（%）/ $C_{avg} * 100$）。

扩增倍数（γ）	IIITH测试集 等错误率 / C_{avg}	LDC测试集 等错误率 / C_{avg}	KGP测试集 等错误率 / C_{avg}
0.5	9.18 / 10.92	39.43 / 43.79	37.56 / 34.89
1	8.75 / 10.01	39.61 / 43.96	38.65 / 36.70
2	8.00 / 9.14	38.19 / 42.71	38.88 / 35.52
3	7.68 / 9.28	38.64 / 42.40	36.80 / 32.92
4	8.03 / 9.68	38.89 / 43.20	35.73 / 34.30

相关失配可以通过基于语音增强的增强方法加以抑制。环境增强有助于减少录音室环境方面的语料库失配。编解码器编码增强带来的语言识别性能提升，可以用不同语料库的中间数据处理和存储协议可能存在差异这一事实来解释。

6.3. 不同增强扩增倍数的语言识别性能

我们通过实验优化增强折叠因子（γ），该因子控制合并后的增强训练集中增强训练语句与原始训练语句的比例。我们使用 $\gamma = \{0.5、1、2、3、4\}$ 训练了独立的语言识别模型，并在表7中比较了它们的语言识别性能。结果表明，随着 γ 从0.5增至3，总体而言，同语料库和跨语料库语言识别性能均有所提升。对于使用IIITH训练的语言识别模型，当 $\gamma = 3$时，同语料库等错误率最低，为7.68%。对于相同的训练-测试组合，当 $\gamma = 2$时，C_{avg} 最低，为9.14。对于使用LDC测试集进行的跨语料库评估，我们观察到，$\gamma = 2$时等错误率最低。而对于KGP测试集，$\gamma = 3$时C_{avg} 最低。这与以下事实相呼应：由于增强带来了更多训练数据，基于时延神经网络的模型性能会随之提升 [82]。但在几乎所有情况下，超过 $\gamma = 3$后，语言识别性能便开始下降。这一有趣的现象可以这样解释：超过 $\gamma = 3$后，与原始训练数据相比，声学多样化训练样本过多。因此，语言识别模型可能开始学习非语言多样性，而非原始数据中实际存在的语言判别线索 [83]。此外，采用较小的扩增倍数 $\gamma = 2$（而非3）也能减少计算量和模型训练时间。

在后续实验中，我们使用 $\gamma = 2$ 从不同信号级增强类别中采样话语。该采样增强集在表10中记为“信号级（$\gamma = 2$）”，在表 11 中记为“增强（$\gamma = 2$）”。在我们的研究中，采样增强集的 $\gamma = 2$ 或 3 在语言识别性能方面表现最佳。此外，对于 $\gamma = 2$，我们还需要降低计算成本。未来有必要验证不同基于语音的分类任务中的最佳增强折叠因子。

6.4. 信号级与特征级级联增强的语言识别性能

我们在信号级增强语句的特征提取步骤中进行特征级增强，以实现级联数据增强。我们选择全部七种信号级增强类别，并使用 $\gamma = 2$ 对其进行采样。对于特征级增强，我们分别评估 SpecAug 和 Mixup。相应的语言识别性能（以 IIITH 作为训练数据）见表8。

表8：采用 IIITH 作为训练语料库，通过级联信号级和特征级增强获得的语言识别性能（等错误率（%）/ $C_{avg} * 100$）。

增强	IIITH测试集 等错误率 / C_{avg}	LDC测试集 等错误率 / C_{avg}	KGP测试集 等错误率 / C_{avg}
信号级 ($\gamma = 2$)	8.00 / 9.14	38.19 / 42.71	38.88 / 35.52
信号级后接 SpecAug	9.82 / 11.20	48.67 / 45.16	39.74 / 40.24
信号级后接 Mixup	9.26 / 10.93	39.70 / 40.61	45.01 / 38.94
多样性感知级联	9.13 / 10.74	39.70 / 43.16	35.17 / 29.32
信号级增强后接级联	8.61 / 10.22	39.16 / 43.04	37.75 / 35.84

分析表8可知，将不同的信号级和特征级增强类别级联，并不能提升语言识别的泛化能力。信号级与特征级增强后的话语可能过于多样化，导致语言分类任务变得困难。因此，在第6.5节中，我们报告了采用多种信号级增强类别进行多样性感知级联增强的实验结果。

6.5. 语言识别性能：多样性感知级联增强

为了以更贴近现实的方式增强话语，我们提出将不同的信号级增强类别进行级联，从而创建更多样化的训练语句。为了有效实现多样化，我们找出了适合级联、两两之间多样性最大的增强类别。

6.5.1. 用于衡量域多样性的指标

对于每个信号级增强类别 A^k，我们都会提取相应话语的嵌入（Z^k）。提取嵌入时，我们使用在组合增强 IIITH 训练集（含 $\gamma = 2$）上训练的语言识别模型。我们还提取了 IIITH 原始未增强话语的嵌入，并将其记为 Z^0。接下来，我们使用三种不同的度量方法测量 Z^0 与 $Z^k \ \forall \ k \in [1,7]$ 之间的嵌入距离：余弦距离（D_{cos}）、KL散度（D_{KL}）和对称KL散度（D_{KL-s}）。对于每种度量方法，我们都计算了嵌入距离与相应验证集 C_{avg}（主要评估指标）分数之间的相关性，并将结果列于表9。由于嵌入距离不一定服从高斯分布，我们使用斯皮尔曼相关系数进行相关性计算。最终，我们选择 KL散度来衡量域多样性距离，因为它与 C_{avg} 分数呈现最强的负相关。原始数据与某个增强类别之间的 KL 散度越大，预期多样性就越高，最终相应的 C_{avg} 也越低。此外，KL 散度的非交换特性（D_{KL-s}不具备这一特性）也支持级联增强的基本假设：$A^{\{i,j\}}$，$A^{\{j,i\}}$。

表9：不同的域多样性度量及其与IIITH验证数据上 C_{avg} 的相关性。

	A^1	A^2	A^3	A^4	A^5	A^6	A^7	相关性
$\mathbf{D}_{cos}$	0.127	0.106	0.221	0.123	0.122	0.009	0.008	-0.107
$\mathbf{D}_{KL}$	0.353	0.133	0.257	0.188	0.205	0.171	0.065	-0.428
D_{KL-s}	0.303	0.177	0.252	0.211	0.232	0.204	0.128	-0.428
C_{avg}	7.11	11.30	11.07	7.82	7.98	7.20	9.95	-

6.5.2. 语言识别性能：最大多样性感知级联增强

接下来，我们展示所提出的级联增强方法的语言识别性能。本实验旨在通过级联彼此差异最大的信号级增强类别，以更切合实际的方式提高训练数据的多样性。考虑所有 $D^{\{i,j\}}{}_{KL}$ 值，对于每个 A^i，我们将其话语与另一个增强类别 A^j 的话语进行级联，使得 $D^{\{i,j\}}{}_{KL}$ 最大。表10展示了 IIITH 语料库中所有增强类别对之间的成对散度。我们为每个增强类别创建一个级联增强集合。随后，我们对这些级联集合进行抽样，并通过 $\gamma = 2$ 将其与原始未增强话语组合起来，训练语言识别模型。此外，我们还使用原始数据、信号级增强数据（$\gamma = 1$）和级联增强数据（$\gamma = 1$）的组合话语训练另一个语言识别模型。表8最后两行展示了使用多样性最大化的级联增强话语时的语言识别性能。

表10：IIITH 语料库中所探索的信号级增强类别的嵌入之间的成对 KL散度。

	A^1	A^2	A^3	A^4	A^5	A^6	A^7
A^1	0.00	0.311	0.385	0.458	**0.713**	0.169	0.312
A^2	**0.313**	0.00	0.271	0.151	0.178	0.228	0.101
A^3	0.518	0.332	0.00	0.210	**0.609**	0.391	0.309
A^4	**0.341**	0.155	0.191	0.00	0.287	0.278	0.194
A^5	0.529	0.173	**0.532**	0.329	0.00	0.340	0.251
A^6	0.187	0.191	0.296	0.311	**0.393**	0.00	0.180
A^7	0.257	0.104	**0.280**	0.174	0.210	0.195	0.00

所提出的最大多样性感知级联增强展现出良好的语言识别性能。有趣的是，对于这种级联增强，其跨语料库性能的相对提升甚至超过了相应的同语料库性能提升。可以继续级联两种以上的增强方法，以在训练数据中引入更多多样性。然而，这将增加计算复杂度。此外，在数据中引入更多多样性可能会扭曲音频数据中的语言特定线索。

我们还测量了级联增强嵌入与未增强嵌入之间的 KL散度；$D^{\{0,\text{cascaded}\}}{}_{KL} = 0.286$。将其与表9中 $k \in [1,7]$ 对应的 $D^{\{0,k\}}{}_{KL}$ 值进行比较，我们发现，在七个增强类别中的六个类别中，$D^{\{0,\text{cascaded}\}}{}_{KL}$ 更高。在本研究中，对于表10的每个 A^i 行，如果列 A^p 显示最高值，我们便将这两个增强类别级联为 $A^{\{i,p\}}$。我们也可以级联总体 $D^{\{i,j\}}{}_{KL}$ 值最高的两个增强类别。然而，这样做无法保证所有探索的增强类别 $A^k \; \forall k = [1,7]$ 的多样性都能通过级联得到保留。本研究的目标是提升在多个完全独立语料库上的泛化能力。所提出的级联算法通过确保音频来自所有探索的增强类别来融入多样性，因此在这类未见评估场景中可能更有益。

6.6. 语言识别性能：域泛化视角

我们按照第5节所述，采用三种领域泛化技术进行了实验：对抗式、多任务处理和最大均值差异。所有这些方法都扩展了当前最先进的ECAPA-TDNN

表 11：所提出的域泛化方法的语言识别性能（等错误率（%）/$C_{avg}*100$）。作为比较，还给出了通过采样（使用 $\gamma = 2$）生成的信号级增强语句的语言识别性能。

方法	训练语料库：IIITH			训练语料库：LDC			训练语料库：KGP		
	IIITH测试集	LDC测试集	KGP测试集	IIITH测试集	LDC测试集	KGP测试集	IIITH测试集	LDC测试集	KGP测试集
	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}	等错误率 / C_{avg}
增强（$\gamma = 2$）	8.00 / 9.14	38.19 / 42.71	38.88 / 35.52	32.91 / 32.53	20.10 / 24.22	48.25 / 41.53	33.67 / 33.79	47.25 / 45.49	8.94 / 7.26
对抗式	9.64 / 10.88	35.40 / 40.68	28.28 / 22.92	34.98 / 36.77	20.80 / 25.14	44.94 / 43.59	36.76 / 34.35	48.26 / 42.73	9.24 / 8.98
多任务处理	7.97 / 9.03	36.29 / 41.89	29.47 / 25.69	36.32 / 36.40	21.28 / 25.47	44.27 / 44.50	38.97 / 36.12	48.12 / 42.59	8.79 / 8.70
MMD	9.06 / 9.78	39.58 / 45.30	33.15 / 26.28	38.89 / 35.81	22.10 / 26.02	46.32 / 45.88	40.58 / 38.15	48.44 / 45.05	11.98 / 13.70
MD-MMD	9.61 / 10.99	36.45 / 41.68	32.56 / 32.68	38.61 / 32.41	21.07 / 24.46	39.32 / 39.94	40.63 / 34.73	48.47 / 44.46	11.71 / 12.00

架构，并通过权衡参数 λ联合优化多个损失函数。按照 [1], ，我们通过实验设置 λ ，以便更好地实现训练收敛。对于对抗方法，最初的15个训练轮次中，我们将 $\lambda = 0$设为0.001。对于另外两种方法，前五个训练轮次中λ 的初始值为0.1。保持 λ 初始值后，我们对所有领域泛化方法逐轮将 λ 增加0.01。表11展示了不同域泛化方法的语言识别性能对比。为进行比较评估，我们还在该表中报告了通过从不同信号级增强类别中采样话语（使用 $\gamma = 2$）生成的增强数据的语言识别性能。

我们观察到一种趋势：在同语料条件下，多任务方法的性能最佳。在跨语料库评估中，对抗模型则最为有效。同语料库评估话语与训练数据之间的域相似性应该更高。因此，领域感知语言学习所学到的伪域带来的额外见解，能够提升此类情况下的语言识别性能。而跨语料库的域可能更加多样，在这种情况下，融入领域不变性会更有帮助。对于使用 IIITH 数据训练的模型，其在 KGP 测试数据上的性能提升高于在 LDC 测试数据上的提升。元信息表明，与 LDC 数据相比，IIITH 和 KGP 数据在多个方面较为相似，这可以解释这一现象。与传统的最大均值差异（MMD）相比，所提出的 MD-MMD 带来了显著的性能提升，尤其是在跨语料库场景中。MD-MMD 方法分别处理每个伪域，并根据多样性距离为其赋予权重。它能提升同语料库和跨语料库的语言识别性能。基于最大均值差异的方法旨在减少源数据（此处为未增强的原始训练数据）与目标数据（此处为包含 $\gamma = 2$ 的增强数据）之间的分布失配。现有研究指出，在着力减少源—目标分布失配的同时，基于最大均值差异的方法会降低类间可分性 [98, 12]。这或许可以解释为什么两种基于最大均值差异的方法在大多数情况下都不如基于对抗式或多任务学习的域泛化方法。

所探索的域泛化方法在同语料库和跨语料库评估中均明显优于基线系统。以使用 IIITH 训练的系统为例，与基线（见表 5）相比，对于 LDC测试集话语，C_{avg} 提升了 5.55%。对于 KGP 测试话语，提升幅度为 6.99%。与 ECAPA-TDNN 模型相比，在 LDC 和 KGP 测试语料库上，C_{avg} 的相对提升幅度分别为 4.14% 和 9.70%。在大多数情况下，域泛化方法的表现也优于所探索的领域多样化方法。域泛化方法将增强训练数据用作伪域。因此，我们将其与使用增强训练集 ($\gamma = 2$) 训练的 ECAPA-TDNN 模型进行比较。比较结果显示，在 LDC 和 KGP 测试语料库上，C_{avg} 的相对提升幅度分别为 2.03% 和 12.60%。如果考察我们所进行的不同实验中的语言识别性能，可以观察到一个明显趋势：每当使用 LDC/KGP 训练的语言识别模型在 KGP/LDC 测试话

语上进行评估时，其性能始终低于其他跨语料库训练-测试组合。表 1 所示的元信息比较表明，这两个语料库之间的失配程度明显更高。例如，KGP 是规模最小的数据库，其话语采自广播新闻录音；而 LDC 是规模最大的数据库，包含电话会话语音（CTS）数据。LDC话语受地区方言影响较大，且录制于多种背景环境中。因此，要在如此迥异的语料库对之间实现跨语料库泛化颇具挑战。针对特定语言开展跨语料库研究，可以为解决这一难题提供更多见解。尽管如此，与基线相比，我们在 LDC测试集和 KGP测试数据上分别取得了 4.55% 和 6.28% 的最佳等错误率提升。另一方面，在同语料库和跨语料库两种情形下，IIITH 话语的性能均持续提升。IIITH 同时包含广播新闻和 CTS 数据，规模也较大。我们认为，该语料库的元特征与另外两个语料库存在一定共性。因此，在 IIITH 语料库上实现跨语料库泛化相对容易。总体而言，考虑到我们使用的语料库资源有限，且由不同机构独立构建、元数据特征差异显著，我们所取得的语言识别性能提升令人鼓舞。

7. 进一步分析与讨论

本节将对所提出的方法在另一个未见语料库上进行盲式跨语料库评估，并提供补充分析和消融研究。我们还将讨论当前研究的若干前景。

7.1. 未见语料库上的评估

迄今为止，我们已在三个语料库上进行了大量实验，并在跨语料库实验中取得了显著改进。这确实论证了机器学习算法理想情况下不应利用任何来自评估集的反馈来优化其配置，而应仅依据验证集进行优化。然而，有人可能会认为，所提出方法的配置和参数是针对选定的三个语料库优化的。因此，为了进一步确保所提出方法的泛化能力，我们在第四个语料库 VoxLingua107（VoxLingua） [89], 上进行了实验，该语料库在系统开发过程中完全未曾见过。VoxLingua 是一个规模庞大且内容多样的语料库，其中的话语是通过在线视频流媒体平台从真实来源自动收集的。我们仅使用该语料库进行了最终评估，从五种语言中各随机选取 500 条时长为 3 秒的话语。这很可能包含音频质量各异、带有各种背景噪声的语音话语，能够代表真实场景中的评估。

在此次盲测评估中，我们考察了在 IIITH 语料库上训练的语言识别模型。图 4 展示了检测误差权衡（DET）曲线，以及根据前文讨论的不同实验得出的同语料库和跨语料库评估的等错误率（EER）。我们发现，所有探索的方法都显著优于基线。与 ECAPA-TDNN 模型在盲测跨语料库评估中的表现相比，领域多样化方法和域泛化方法分别使 EER 相对降低了 4.40% 和 5.23%。在领域多样化实验中，所提出的多样性感知级联增强在盲测评估中表现最佳；而对抗学习方法在此次评估中优于所有其他域泛化方法。

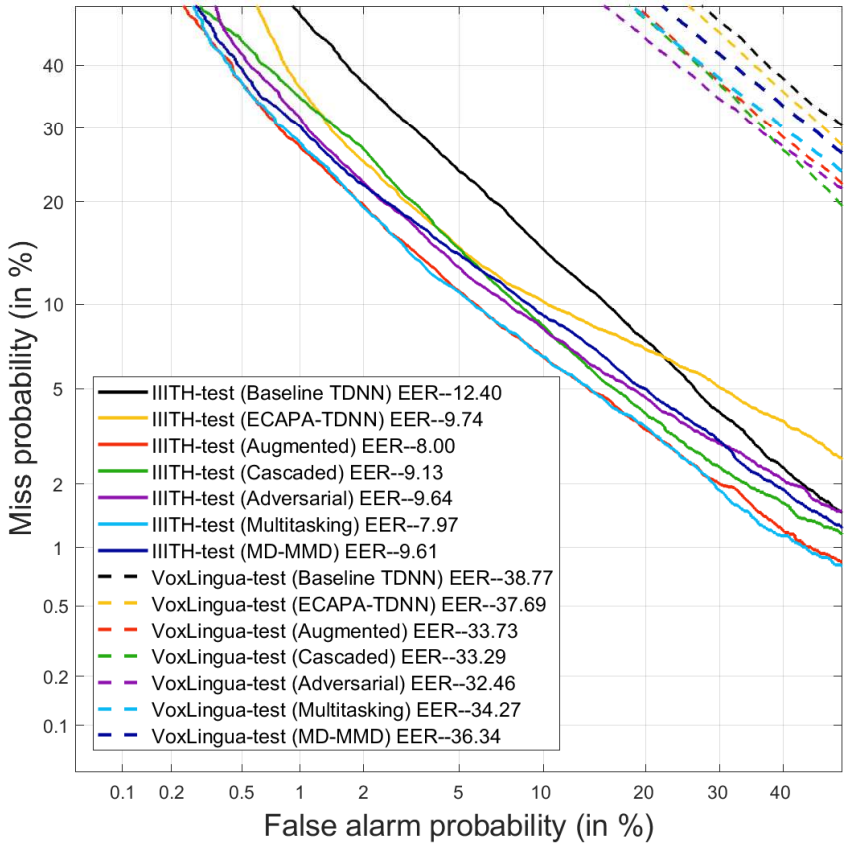

图 4. 基线方法和所提出方法在同语料库（IIITH测试集）及未见跨语料库（VoxLingua-test）条件下的检测错误权衡图。

7.2. 评估领域多样化

我们提出了最大多样性感知级联增强，以有效扩充训练数据，从而提升跨语料库泛化能力。我们还将分析此类增强方法是否提升了域多样性。本研究采用的数据驱动型域多样性度量，是基于使用增强数据训练的语言识别模型所提取的嵌入。现在，我们使用一个在不同音频数据上训练、且独立于语言识别的模型来开展这项多样性分析。我们选用一个为声学场景分类任务预训练的模型。具体而言，我们选择 OpenL3 嵌入，来探究三个语料库之间的背景与环境差异 [33]。

图 5 可视化了三个数据库中未增强训练集的嵌入，分别展示了排除和包含 IIITH增强训练集的情况。我们观察到，由未增强的原始训练数据组成的三个语料库（图 5 (a)）在 OpenL3 嵌入空间中有一定程度的分离，这表明存在显著的非语言差异。我们还观察到，IIITH 增强版话语的 OpenL3 嵌入（图 5 (b)）均匀分布于三个语料库中，表明所提出的方法能够有效实现领域多样化。

我们还在图 6 中分析了三个数据库的整体长期平均频谱（LTAS），并将其与增强后的 IIITH 长期平均频谱进行了比较。增强后的频谱与 KGP-LTAS 频谱更为接近，表明跨语料库性能有所提升。我们还使用孟加拉语和旁遮普语这两种语言，对比了特定语言的长期平均频谱。这两张图表明，跨语料库失配在不同语言间并不均匀。这项工作未来有望进一步拓展，对特定语言的跨语料库性能和泛化能力进行分析。

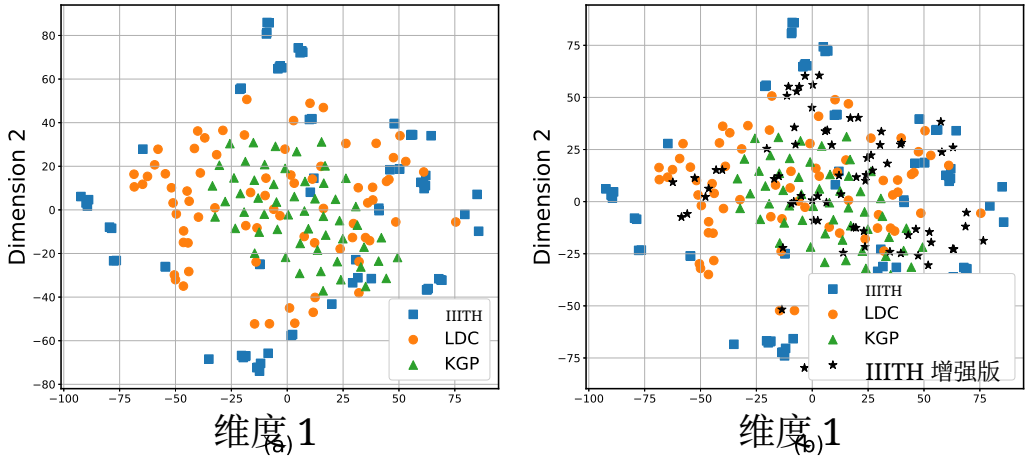

图 5. IIITH、LDC 和 KGP 语料库的 OpenL3 嵌入的 t-SNE 可视化：(a) 训练集；(b) 进一步加入 IIITH 增强版（$\gamma = 2$）数据集。

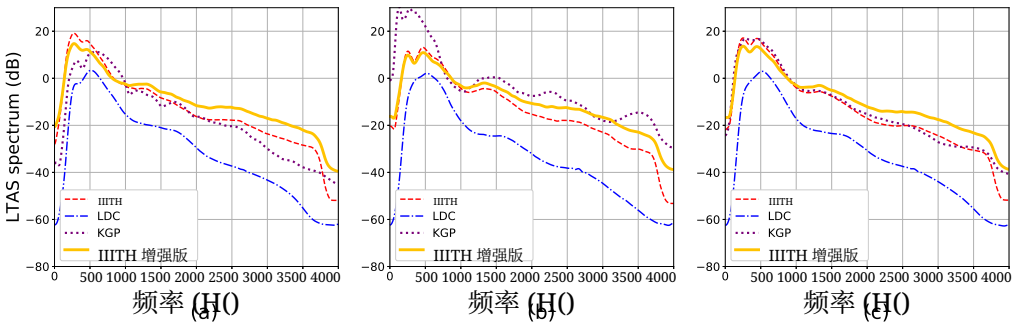

图 6. 三个语料库与 IIITH 增强数据（含 $\gamma = 2$）的长期平均频谱（LTAS）比较：（a）孟加拉语；（b）旁遮普语；（c）所有训练话语。

7.3. 领域多样化中的语言多样性与非语言多样性

在第4.2节中，我们讨论扩增倍数的影响，即训练集中增强语句数量与未增强原始训练话语数量之比。在这种情况下，我们遵循文献 [81, 25] 中讨论的增强文件选择惯例。这包括为每种方法生成增强副本，随后

表12：增强训练集中内容多样性与非语言多样性对同语料库和跨语料库语言识别性能（等错误率（%）/ $C_{avg} * 100$）的影响。此处使用 IIITH 语料库创建了三个增强集，每个增强集的内容多样性和非语言多样性程度各不相同。

增强	IIITH 等错误率/ C_{avg}	LDC 等错误率/ C_{avg}	KGP 等错误率/ C_{avg}
第一种情景	8.75 / 10.01	39.61 / 43.96	38.65 / 36.70
第二种情景	7.40 / 9.41	42.31 / 45.99	28.00 / 29.75
第三种情景	7.00 / 8.98	46.15 / 41.55	31.00 / 30.37

从汇总数据中随机选择增强文件，以满足所需的扩增倍数。由于从汇总数据中进行随机采样，同一话语可能拥有多个增强副本，即相同内容会以不同的非语言多样性重复多次。而有些话语可能完全没有任何增强副本。在本节中，我们进一步探究增强数据中内容与非语言多样性各自的影响。现有音频增强文献对这类研究的探讨并不广泛，尤其是在说话人和语言识别问题方面。因此，我们认为纳入此类讨论对于数据增强研究社群十分重要。

我们采用三种情景创建三个增强集（含 $\gamma = 1$）。在第一种情景中，我们遵循通常的做法，对增强数据进行采样；本节开头讨论了这一做法，第6.3节也采用了这一做法。如上所述，在这种情景下，某些内容可能会重复，而某些话语则可能没有增强副本。在第二种情景中，我们施加一项限制：从同一音频文件的所有增强版本中，随机选取一个版本。请注意，这两种情景包含的音频文件数量相同（由于 $\gamma = 1$，是原始数据中文件数量的两倍），但前一种情景的内容可变性低于后一种。我们还考虑了第三种情景，进一步限制内容可变性：随机选取 1/2 的原始话语。随后，对于每个选中的话语，我们随机选取任意三个增强副本（来自七种不同的信号级增强类别）。第三个增强集包含的音频文件总数也与另外两种情景相同，但其内容可变性更低，非语言可变性更高。

表12展示了以 IIITH 为训练语料库时，三种数据增强情景下的语言识别结果。由于音频文件选择具有随机性，我们将每项实验进行了五次，并报告了平均性能。表12中的结果表明，对内容重复施加限制有助于提升两种测试条件（IIITH 和 KGP）下的语言识别性能。元信息显示，IIITH 和 KGP 都主要是广播新闻数据。因此，这些数据库中的发音方式、口音和方言均保持标准。相比之下，LDC 属于会话数据，应包含多种口音和方言变体。对于在评估中使用的此类数据库，在数据增强期间限制内容重复并无帮助。我们认为，在领域多样化的背景下，进一步探索内容多样性与非语言多样性是一个值得开展的未来研究方向。

7.4. 域不变方法与域感知方法

所探索的域泛化方法的语言识别性能呈现出一种趋势：对抗模型适用于跨语料库泛化，而多任务学习模型有助于提升同语料库语言识别性能。因此，在语言训练期间处理域信息时，我们有两种显然相反的方法：域不变

和域感知 [1]。与跨语料库话语相比，同语料库数据更有可能与所创建的伪域具有较高的领域相似性。因此，多任务学习架构中额外的伪域分类任务，可能会使语言识别模型意识到同语料库测试话语中可能存在的多样性。相反，跨语料库评估语句所属的域可能与伪域相似性较低。在这种情况下，域不变方法通过降低模型对领域多样性的敏感度，有助于提升跨语料库语言识别性能。未来的工作可以考虑探索一种算法，以便针对给定的测试音频 [1]动态选择域感知或域不变模型。

7.5. 跨语料库性能差距：尚存挑战与解决方案

本文提出的方法在跨语料库评估中取得了显著提升。然而，同语料库评估与跨语料库评估之间的性能差距仍然相当大。这可能有几个原因：(i) 用于领域多样化的增强方法无法准确模拟现实世界的域，而音频数据模拟本身也是一项具有挑战性的任务 [71]。本研究所提出的域泛化问题也依赖于增强数据。我们推测，条件更贴近现实的增强数据有助于缩小性能差距。(ii) 由于问题设定的限制，我们不使用任何目标域数据，这限制了模型针对某个特定域进行优化的空间。(iii) 在跨语料库条件下表现良好的域不变方法，并未能彻底从潜在表示中去除与域相关的特征。不过，当前研究的主要动机，是评估在具有挑战性的跨语料库条件下，领域多样化和域泛化的能力上限，并着重考察相较于现有基线的相对提升。参照短时长的美国国家标准与技术研究院 LRE 评估条件 [74], ，我们也采用了时长仅为 3 秒的测试语音，这使语言识别问题更加困难 [30]。在 LRE17 的 3 秒时长条件下，当前最先进的语言识别系统的等错误率约为 15%；在该条件下，系统的训练和测试均使用电话会话语音 [62]。尽管等错误率如此之高的系统并不可靠，我们仍相信，所取得的相对提升为解决这一具有挑战性的任务指明了一个有希望的方向：对短时长测试语音进行跨语料库泛化。

在当前研究中，所提出的领域多样化和领域泛化方法仅应用于 ECAPA-TDNN，这是一种用于语言分类的特定模型，其架构通过判别式方式进行训练。该研究可以扩展至其他判别式和生成式方法。近年来，自监督学习（SSL） [56] 已广泛应用于各种任务，包括语言识别 [46], ，后者利用来自其他源的大量未标注音频数据。我们认为，使用自监督学习框架进行跨语料库评估，并评估其与同语料库评估之间的性能差距，是值得探索的未来研究方向。

8. 结论

本研究探讨了领域多样化和域泛化在提升低资源语言识别任务中的跨语料库泛化能力方面是否适用。我们探讨了用于领域多样化的现有及新型音频数据增强方法。对于域泛化，我们研究了对抗学习、多任务学习和MMD损失，以及其结合伪域的新变体。除了在三个广泛使用的

印度语言识别语料库上开展大量实验外，我们还在一个未见语料库上进行了盲测评估，以进一步验证所提方法的有效性。我们将本研究的主要发现总结如下：

- 不同的评估语料库可能具有多样的非语言学特征。因此，除了传统的音频增强之外，探索扰动音频数据的新来源能够进一步提高训练数据的多样性，从而有效提升未见跨语料库泛化能力。例如，对于基于所提出的语音增强或编解码器的增强方法，最有望提升在 LDC 和 KGP 测试数据上的鲁棒性，适用于使用 IIITH 数据训练的语言识别模型。
- 应用领域多样化的根本动机，是通过模拟与潜在未见域相关的非语言学特征来减少跨语料库失配。因此，以更贴近现实的方式增强音频数据至关重要。为此，我们提出了采用优化的折叠因子的最大多样性感知级联增强；与传统增强方法相比，它是在盲测评估中更有效的解决方案。

- 我们的目标是在训练数据有限的情况下，实现语言识别模型在多种未见域上的更好泛化。因此，我们不采用传统领域自适应方法来针对特定目标域提升鲁棒性，而是选择域泛化。与领域多样化方法相比，该方法将不同的增强类别用作伪域，进一步提升语言识别性能。
- 我们通过域不变方法和域感知方法，以两种本质上截然相反的方式探索域泛化。采用梯度反转或基于最大均值差异的方法训练域不变语言识别模型，使其对非语言的域特定失配不那么敏感。在域感知的情形下，我们则通过多任务学习更新语言识别模型，使其额外学习域特定线索。研究发现，域不变学习更有利于跨语料库泛化，而域感知则能提升同语料库语言识别性能。

尽管所提出的领域多样化和域泛化显著提升了跨语料库语言识别性能，但仍需进一步研究，以缩小同语料库与跨语料库性能之间的差距。探索其他先进的数据增强方法，例如采用语音转换、文本到语音合成和风格迁移，或许大有裨益。我们从整个数据库的角度处理跨语料库问题。以特定语言为单位研究跨语料库语言识别，或能进一步揭示这一问题的内在规律。我们对语音内容与非语言可变性的进一步研究表明，这是另一个值得探索的方向。

致谢

第一作者还要感谢美国宾夕法尼亚大学语言数据联盟（LDC）颁发的 LDC数据奖学金。我们还要感谢卡拉格普尔印度理工学院的 KS Rao 教授和海得拉巴国际信息技术学院的 AK Vuppala 教授，他们为 KGP 和 IIITH数据集提供了帮助。最后，作者谨此感谢副编辑和审稿人认真审阅本文、提出详尽意见和建设性建议，这些意见和建议显著提升了稿件的内容。

参考文献

[1] Adi, Y., Zeghidour, N., Collobert, R., Usunier, N., Liptchinsky, V., Synnaeve, G., 2019年。是否反转梯度：语音识别中对抗式学习与多任务学习的实证比较，载于：ICASSP，IEEE。第 3742–3746 页。[2] Alum¨ae, T., Kukk, K., 2022年。口语语言识别的预训练方法：塔尔图科技大学提交 OLR 2021 挑战赛的方案，载于：Odyssey：说话人和语言识别研讨会，ISCA。第 240–247 页。[3] Benyassine, A., Shlomot, E., Su, H., Massaloux, D., Lamblin, C., Petit, J., 1997年。用于数字语音与数据同时传输应用的静音压缩方案（建议 G.729 附录 B），针对 G.729 进行了优化。IEEE 通信杂志 35，64–73。[4] Berouti, M., Schwartz, R., Makhoul, J., 1979年。受声学噪声污染的语音增强，载于：ICASSP，IEEE。第 208–211 页。[5] Beyan, C., Shahid, M., Murino, V., 2021年。RealVAD：一个现实世界数据集，以及一种基于身体运动分析的语音活动检测方法。《IEEE 多媒体汇刊》23，2071–2085。[6] Blanchard, G., Lee, G., Scott, C., 2011年。从若干相关分类任务泛化到新的未标记样本。《神经信息处理系统进展》24，2178–2186。[7] Brookes, M., 等，1997年。Voicebox：用于 MATLAB 的语音处理工具箱。软件，网址： [1997年3月 2011] 起可从 www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox.html 获取。47。[8] Br¨ummer, N., Du Preez, J., 2006年。与应用无关的说话人检测评估。《计算机语音与语言》20，230–275。[9] Caruana, R., 1997年。多任务学习。《机器学习》28，41–75。[10] Cha, J., Chun, S., Lee, K., Cho, Han-Cheol andPark, S., Lee, Y., Park, S., 2021年。SWAD：通过寻求平坦极小值实现域泛化，载于：《神经信息处理系统进展》，第 22405–22418 页。[11] Chakraborty, J., Chakraborty, B., Bhattacharya, U., 2021年。口语语言的 DenseRecognition，载于：国际模式识别会议（ICPR），IEEE。第 9674–9681 页。[12] Chen, Y., Song, S., Li, S., Wu, C., 2019年。一种基于最大均值差异的领域自适应算法图嵌入框架。《IEEE 图像处理汇刊》29，199–213。[13] Chettri, B., Hautam¨aki, R.G., Sahidullah, M., Kinnunen, T., 2021年。数据质量作为语音反欺骗泛化能力的预测指标，载于：INTERSPEECH，ISCA。第 1659–1663 页。[14] Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schl¨ogl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., 等，2019年。人机交互中的语音现状：趋势、主题与挑战。《与计算机交互》31，349–371。[15] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J., 2018年。通过保留自相似性和领域差异性的图像间领域自适应实现行人重识别，载于：CVPR，第 994–1003 页。[16] Desplanques, B., Thienpondt, J., Demuynck, K., 2020年。ECAPA-TDNN：基于时延神经网络的说话人验证中的增强型通道注意力、传播与聚合，载于：INTERSPEECH，ISCA。第 1–5 页。[17] Dey, S., Saha, G., Sahidullah, M., 2021年。跨语料库语言识别：对印度语言的初步研究，载于：EUSIPCO，IEEE。第 546–550 页。[18] Dey, S., Sahidullah, M., Saha, G., 2022年。从机器学习视角概述印度口语语言识别。《ACM 亚洲及低资源语言信息处理汇刊》21，1–45。[19] Ding, Z., Fu, Y., 2017年。带结构化低秩约束的深度域泛化。《IEEE 图像处理汇刊》27，304–313。[20] Doire, C.S., 等，2016年。对受混响和噪声污染的语音进行单通道在线增强。《IEEE/ACM 音频、语音与语言处理汇刊》25，572–587。[21] Du, C., Li, H., Lu, Y., Wang, L., Qian, Y., 2021年。端到端语码转换语音识别的数据增强，载于：口语语言技术研讨会（SLT），IEEE。第 194–200 页。[22] Duroselle, R., Jouvet, D., Illina, I., 2020年。通过度量学习损失函数减少语言识别中 x 向量空间的领域失配，载于：INTERSPEECH，第 447–451 页。[23] Ferrer, L., Castan, D., McLaren, M., Lawson, A., 2022年。一种用于口语语言识别的判别式分层 PLDA 模型。《IEEE/ACM 音频、语音与语言处理汇刊》30，2396–2410。[24] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016年。神经网络的领域对抗训练。《机器学习研究期刊》17，2096–2030。[25] Garcia-Romero, D., Sell, G., Mccree, A., 2020年。MagNetO：用于提升说话人识别性能的 X 向量幅度估计网络及偏移量，载于：Odyssey：说话人和语言识别研讨会，ISCA。第 1–8 页。[26] Gerczuk, M., Amiriparian, S., Ottl, S., Schuller, B., 2021年。EmoNet：用于多语料库语音情感识别的迁移学习框架。《IEEE 情感计算汇刊》，1–1。

[27] Gerkmann, T., Hendriks, R.C., 2011年. 一种低复杂度、低跟踪延迟的无偏基于最小均方误差噪声功率估计方法. IEEE/ACM 音频、语音与语言处理汇刊 20, 1383–1393. [28] Gideon, J., McInnis, M.G., Provost, E.M., 2021年. 通过对抗式判别域泛化（ADDoG）改进跨语料库语音情感识别. IEEE 情感计算汇刊 12, 1055–1068.[29] Gillespie, S., Logan, Y.Y., Moore, E., Laures-Gore, J., Russell, S., Patel, R., 2017年. 用于构音障碍存在与否分类的跨数据库模型，载于：INTERSPEECH，ISCA. 第3127–3131页. [30] Gonzalez-Dominguez, J., 等, 2015年. 使用深度神经网络对短话语进行逐帧语言识别. 神经网络 64, 49–58. [31] Greenberg, C.S., Martin, A.F., Przybocki, M.A., 2012年. 2011年 NIST语言识别评测，载于：INTERSPEECH，ISCA. 第34–37页. [32] Gretton, A., Borgwardt, K., Rasch, M., Sch¨olkopf, B., Smola, A., 2006年. 一种用于双样本问题的核方法. 神经信息处理系统进展 19. [33] Grollmisch, S., Cano, E., Kehling, C., Taenzer, M., 2021年. 分析预训练嵌入在音频分类任务中的潜力，载于：EUSIPCO，IEEE. 第790–794页. [34] Gulrajani, I., Lopez-Paz, D., 2021年. 探寻失落的域泛化，载于：国际学习表征会议. [35] Hu, X., Yang, Y.H., 2017年. 跨数据集与跨文化的音乐情绪预测：以西方和中国流行歌曲为例. IEEE 情感计算汇刊 8, 228–240. [36] Iqbal, T., Helwani, K., Krishnaswamy, A., Wang, W., 2021年. 通过一致性学习增强音频增强方法，载于：ICASSP，IEEE. 第646–650页. [37] Kang, W., Alam, M.J., Fathan, A., 2022年. 面向域不匹配场景的基于深度学习的端到端口语语言识别系统，载于：语言资源与评估会议，第7339–7343页. [38] Karen, J., 等, 2017年. 多语言电话会话语音2011——南亚 LDC2017S14. 网络下载. 费城：语言数据联盟. - . [39] Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A., 2012年. 消除数据集偏差的影响，载于：欧洲计算机视觉会议，Springer. 第158–171页. [40] Korshunov, P., Marcel, S., 2019年. 一项关于语音呈现攻击检测的跨数据库研究，载于：《生物特征反欺骗手册》. Springer，第363–389页. [41] Kumawat, P., Routray, A., 2021年. 应用时延神经网络架构分析语音情感识别中的时长依赖关系，载于：INTERSPEECH，ISCA. 第3410–3414页. [42] Li, H., Ma, B., Lee, K.A., 2013年. 口语语言识别：从基础到实践.《IEEE 会刊》101, 1136–1159.[43] Li, L., Li, Z., Liu, Y., Hong, Q., 2021年. 面向语言识别的深度联合学习. 神经网络 141, 72–86. [44] Li, Z., Zhao, M., Hong, Q., Li, L., Tang, Z., Wang, D., Song, L., Yang, C., 2020年. AP20-OLR 挑战赛：三项任务及其基线方法，载于：亚太信号与信息处理协会年度峰会暨会议（APSIPA ASC），IEEE. 第550–555页. [45] Liu, H., Garcia Perera, L.P., Khong, A., Styles, S., Khudanpur, S., 2022a. PHO-LID：一种融合声学语音学信息与音系音位信息的统一语言识别模型，载于：INTERSPEECH，ISCA. 第2233–2237页. [46] Liu, H., Perera, L.P.G., Khong, A.W., Chng, E.S., Styles, S.J., Khudanpur, S., 2022b. 用于口语语言识别的高效自监督学习表征. IEEE信号处理精选主题期刊 16, 1296–1307. [47] Liu, H., Perera, L.P.G., Khong, A.W., Dauwels, J., Styles, S.J., Khudanpur, S., 2022c. 使用带知识蒸馏的双模态模型增强语言识别，载于：Odyssey：说话人和语言识别研讨会，ISCA. 第248–254页. [48] Long, M., Cao, Y., Wang, J., Jordan, M., 2015年. 使用深度自适应网络学习可迁移特征，载于：国际机器学习会议，PMLR. 第97–105页. [49] Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J., Moreno, P.J., 2016年. 深度前馈神经网络在自动语言识别中的应用.《计算机语音与语言》 40, 46–59. [50] Loshchilov, I., Hutter, F., 2018年. 解耦权重衰减正则化，载于：ICLR. [51] Maity, S., A.K., V., Rao, K., Nandi, D., 2012年. 用于语言识别的 IITKGP-MLILSC 语音数据库，载于：全国通信会议（NCC），IEEE. 第1–5页. [52] Mandava, T., Vuddagiri, R.K., Vydana, H.K., Vuppala, A.K., 2019年. 对基于 LSTM-CTC 的印度语言识别联合声学模型的研究，载于：自动语音识别与理解研讨会（ASRU），IEEE. 第389–396页. [53] Mandava, T., Vuppala, A.K., 2019年. 用于印度语言识别的注意力残差时延神经网络，载于：当代计算国际会议（IC3），IEEE. 第1–5页.

[54] Martinez, D., Plchot, O., Burget, L., Glembek, O., Matˇejka, P., 2011年. i向量空间中的语言识别，载于：INTERSPEECH，ISCA. [55] Mauch, M., Ewert, S., 2013年. 音频退化工具箱及其在鲁棒性评估中的应用，载于：国际音乐信息检索学会会议（ISMIR），库里蒂巴，巴西. [56] Mohamed, A., Lee, H.y., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.W., Livescu, K., Maaløe, L., Sainath, T.N., Watanabe, S., 2022年. 自监督语音表征学习：综述. IEEE信号处理精选主题期刊 16, 1179–1210. [57] Monteiro, J., Alam, J., Falk, T.H., 2019年. 结合注意力特征池化的残差卷积神经网络，用于基于短时语音的端到端语言识别. 《计算机语音与语言》 58, 364–376. [58] Moreno-Torres, J.G., Raeder, T., Alaiz-Rodr´ıguez, R., Chawla, N.V., Herrera, F., 2012年. 对分类中数据集偏移的统一观点. 《模式识别》 45, 521–530. [59] Mozilla, . Common Voice. https://commonvoice.mozilla.org/en. 访问日期：2020年1月12日. [60] Mushtaq, Z., Su, S.F., 2020年. 使用正则化深度卷积神经网络和数据增强进行环境声音分类. 《应用声学》 167, 107389. [61] Nadimpalli, A.V., Rattani, A., 2022年. 改进深度伪造检测器的跨数据集泛化，载于：CVPR，IEEE/CVF. 第91–99页. [62] Padi, B., Mohan, A., Ganapathy, S., 2020年. 探索语言识别中的相关性与序列建模. IEEE/ACM 音频、语音与语言处理汇刊 28, 1223–1232. [63] Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q., 2010年. 通过迁移成分分析实现领域自适应. IEEE神经网络汇刊 22, 199–210. [64] Pandey, A., Wang, D., 2022年. 用于语音增强的自注意力RNN，以提升跨语料库泛化能力. IEEE/ACM 音频、语音与语言处理汇刊 30, 1374–1385. [65] Park, D.S., 等, 2019年. SpecAugment：一种用于自动语音识别的简单数据增强方法，载于：INTERSPEECH，ISCA. 第2613–2617页. [66] Paszke, A., Gross, S., Massa, F., Lerer, A., 2019年. PyTorch：一种命令式风格、高性能的深度学习库，载于：NeurIPS，第8024–8035页. [67] Paul, D., Sahidullah, M., Saha, G., 2017年. 欺骗攻击对策的泛化：以 ASVspoof 2015 和 BTAS 2016 语料库为例，载于：ICASSP，IEEE. 第2047–2051页. [68] Povey, D., 等, 2011年. Kaldi 语音识别工具包，载于：自动语音识别与理解研讨会（ASRU），IEEE. [69] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022年. 通过大规模弱监督实现鲁棒语音识别. OpenAI博客 . [70] Reddy, V.R., Maity, S., Rao, K.S., 2013年. 使用多层级频谱和韵律特征识别印度语言. 《国际语音技术期刊》 16, 489–511. [71] Ribas, D., Vincent, E., Calvo, J.R., 2016年. 真实场景中的语音失真条件及其对语音处理应用的影响研究，载于：口语语言技术研讨会（SLT），第13–20页. [72] Rossenbach, N., Zeyer, A., Schl¨uter, R., Ney, H., 2020年. 为基于注意力的语音识别系统生成合成音频数据，载于：ICASSP，IEEE. 第7069–7073页. [73] Ruder, S., 2017年. 深度神经网络多任务学习概述. arXiv 预印本 arXiv:1706.05098 . [74] Sadjadi, S.O., Kheyrkhah, T., Tong, A., Greenberg, C.S., Reynolds, D.A., Singer, E., Mason, L.P., Hernandez-Cordero, J., 2018年. 2017年 NIST语言识别评测，载于：Odyssey：说话人和语言识别研讨会，ISCA. [75] Salamon, J., Bello, J.P., 2017年. 用于环境声音分类的深度卷积神经网络和数据增强. IEEE 信号处理快报 24, 279–283. [76] Sarfjoo, S., Madikeri, S., Motlicek, P., Marcel, S., 2020年. 使用有限数据对文本无关说话人验证进行有监督领域自适应，载于：INTERSPEECH，第3815–3819页. [77] Schuller, B., Vlasenko, B., Eyben, F., W¨ollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010年. 跨语料库声学情感识别：方差与策略. IEEE 情感计算汇刊 1, 119–131. [78] Shen, P., Lu, X., Li, S., Kawai, H., 2017年. 用于口语语言识别的条件生成对抗网络分类器，载于：INTERSPEECH，ISCA. 第2814–2818页. [79] Singh, P., Saha, G., Sahidullah, M., 2021年. 使用常数Q变换进行非线性频率扭曲，以实现语音情感识别，载于：国际计算机通信与信息学会议（ICCCI），IEEE. 第1–6页. [80] Snyder, D., Chen, G., Povey, D., 2015年. MUSAN：音乐、语音和噪声语料库. arXiv 预印本 arXiv:1510.08484 .[81] Snyder, D., 等, 2018a. 使用 x 向量进行口语语言识别，载于：Odyssey：说话人和语言识别研讨会，第105–111页. [82] Snyder, D., 等, 2018b. X 向量：用于说话人识别的鲁棒DNN嵌入，载于：ICASSP，IEEE. 第5329–5333页.

[83] Sturm, B.L., 2014年。确定音乐信息检索系统是否是“马”的一种简单方法。《IEEE 多媒体汇刊》16, 1636–1644. [84] Tang, Z., Wang, D., Song, L., 2019年。AP19-OLR 挑战赛：三项任务及其基线方法，载于：亚太信号与信息处理协会年度峰会暨会议（APSIPA ASC），IEEE。第 1917–1921 页。[85] Thienpondt, J., Desplanques, B., Demuynck, K., 2022年。通过利用语言信息解决跨语言说话人验证中的分数偏移问题，载于：ICASSP，IEEE。第 7187–7191 页。[86] Toledo-Ronen, O., Sorin, A., 2013年。基于语音的悲伤与愤怒识别及跨语料库评估，载于：ICASSP，IEEE。第 7517–7521 页。[87] Tong, F., Zhao, M., Zhou, J., Lu, H., Li, Z., Li, L., Hong, Q., 2021年。ASV-Subtools：自动说话人验证开源工具包，载于：ICASSP，IEEE。第 6184–6188 页。[88] Tsakalidis, S., Byrne, W., 2005。异构数据源的声学训练：普通话电话会话语音转录实验，载于：ICASSP，IEEE。第 I–461 页。[89] Valk, J., Alum¨ae, T., 2021年。VoxLingua107：口语语言识别数据集，载于：口语语言技术研讨会（SLT），IEEE。第 652–658 页。[90] Vlasenko, B., Philippou-H¨ubner, D., Wendemuth, A., 2013年。跨语料库情绪分类的参数优化问题，载于：Humaine 协会情感计算与智能交互会议，IEEE。第 454–459 页。[91] Vlasenko, B., Prylipko, D., B¨ock, R., Wendemuth, A., 2014。为实际应用创建稳健情绪分类器而对语音模式变异性进行建模。《计算机语音与语言》28, 483–500. [92] Vuddagiri, 等，2018年。用于印度语言识别的 IIITH-ILSC 语音数据库，载于：资源匮乏语言口语技术，IEEE。第 56–60 页。[93] Wang, F., Cheng, J., Liu, W., Liu, H., 2018年。用于人脸验证的加性边距 Softmax。IEEE 信号处理快报 25, 926–930. [94] Wang, J., Zhu, X., Gong, S., Li, W., 2018年。用于无监督行人重识别的可迁移联合属性-身份深度学习，载于：CVPR，IEEE。第 2275–2284 页。[95] Wang, M., Deng, W., 2018年。深度视觉域自适应：综述。《神经计算》312, 135–153. [96] Wei, S., Zou, S., Liao, F., 等，2020年。基于深度学习的音频分类数据增强方法比较，载于：《物理学杂志：会议系列》，IOP 出版社。[97] Xia, W., Zhang, C., Weng, C., Yu, M., Yu, D., 2021年。使用原型动量对比学习进行自监督文本无关说话人验证，载于：ICASSP，IEEE。第 6723–6727 页。[98] Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W., 2017年。谨防类别权重偏差：用于无监督域自适应的加权最大均值差异，载于：CVPR，IEEE。第 2272–2281 页。[99] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018年。mixup：超越经验风险最小化，载于：ICLR。[100] Zhang, Y., Deng, B., Tang, H., Zhang, L., Jia, K., 2020年。无监督多类别域自适应：理论、算法与实践。IEEE 模式分析与机器智能汇刊。[101] Zhang, Y., Yang, Q., 2021年。多任务学习综述。IEEE 知识与数据工程汇刊，1–1。[102] Zhang, Z., Weninger, F., W¨ollmer, M., Schuller, B., 2011年。跨语料库声学情感识别中的无监督学习，载于：自动语音识别与理解研讨会（ASRU），IEEE。第 523–528 页。[103] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C., 2022年。域泛化：综述。IEEE 模式分析与机器智能汇刊。[104] Zhu, R., Zhang, T., Zhao, Q., Wu, Z., 2015年。用于跨数据库面部表情识别的迁移学习方法，载于：国际生物特征识别会议（ICB），IEEE。第 293–298 页。[105] Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., Xiong, H., He, Q., 2020年。用于图像分类的深度子域自适应网络。IEEE 神经网络与学习系统汇刊 32, 1713–1722. [106] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q., 2020年。迁移学习综述。《IEEE 会刊》109, 43–76。