计算机科学
COMPUTER SCIENCE

低资源语言自动语音识别中的数据处理与数据增强综述

杨健, 孙浏, 张丽芳

引用本文

杨健, 孙浏, 张丽芳. 低资源语言自动语音识别中的数据处理与数据增强综述[J]. 计算机科学, 2025, 52(8): 86-99.

YANG Jian, SUN Liu, ZHANG Lifang. Survey on Data Processing and Data Augmentation in Low-resource Language Automatic Speech Recognition [J]. Computer Science, 2025, 52(8): 86-99.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900009

低资源语言自动语音识别中的数据处理与数据增强综述

杨　健[1,2]　孙　浏[1,2]　张丽芳[1]
1 玉溪师范学院工学院　云南 玉溪 653100
2 玉溪师范学院云南省智慧城市网络空间安全重点实验室　云南 玉溪 653100

摘　要　由于标注语音数据不足,端到端自动语音识别(Automatic Speech Recognition,ASR)技术难以直接应用到低资源语言场景,低资源语言 ASR 也成为 NLP 领域的热点问题。目前,低资源环境下 ASR 的研究可以从数据增强和模型改进两方面开展,以低资源语言 ASR 中的训练数据处理为主要研究对象,重点从数据增强、样本处理、特征工程等角度,对近年来该领域的重要研究成果进行梳理和总结。分析了不同类型的数据增强方案,强调未配对语音和文本的利用,并从特征抽取、嵌入和融合等不同方面对低资源环境下 ASR 的特征工程进行分析和总结,阐述了低资源语音语料库建设等问题,并对低资源环境下用于语音识别的数据增强技术未来可以进一步深入研究的重要方向进行展望。

关键词:低资源;自动语音识别;数据增强;特征表示
中图分类号　TP391

Survey on Data Processing and Data Augmentation in Low-resource Language Automatic Speech Recognition

YANG Jian[1,2],SUN Liu[1,2] and ZHANG Lifang[1]
1 College of Engineering,Yuxi Normal University,Yuxi,Yunnan 653100,China
2 Yunnan Provincial Key Laboratory of Cyberspace Security for Smart Cities,Yuxi Normal University,Yuxi,Yunnan 653100,China

Abstract　Due to the absence of transcribed speech,applying end-to-end ASR technology to low-resource language is challenging,making low-resource language ASR is a prominent research topic in NLP. Research on ASR in low-resource settings can be approached from two main aspects:data augmentation and model improvement. This paper focuses on the processing of training data in low-resource language ASR and summarizes the important research results in this field in recent years from the perspectives of data augmentation,sample processing,and feature engineering. Different types of data augmentation schemes are analyzed,and the utilization of unpaired speech and unpaired text is elaborated in detail. The feature engineering of ASR in low-resource scenarios is analyzed and summarized from different aspects such as feature extraction,embedding,and fusion. Finally,additional issues such as the construction of low-resource speech corpora are elaborated,and important directions for further research in low-resource language ASR are prospected.

Keywords　Low-resource,Automatic speech recognition,Data augmentation,Feature representation

1　引言

随着深度学习的快速发展,自动语音识别(ASR)理论研究取得巨大进步,实际应用也早已达到商业化程度,然而这是对资源丰富的语言而言。据文献研究,目前全世界有 5000～7000 种语言,其中大部分是低资源语言[1-2]。商业化的 ASR 应用难以覆盖这些语言,因此低资源语音识别也是 NLP 中最重要的开放问题[3]。基于深度架构的端到端(E2E)ASR 模型在 NLP 的多数下游任务上都取得了最好结果,然而,它倾向于使用海量训练数据以获得较高精度,在低资源环境下,传统端到端模型的准确率反而低于混合模型[4]。由于低资源的基本特征,数据增强对于端到端深度模型下的低资源语言 ASR 具有重要意义。

1.1　低资源的定义

对于低资源语音识别中“低”的界限定义,学术界是有争议的[5],文献[6]认为非常低资源、低资源和中等资源的语音分别为 10 分钟、1 小时和 10～50 小时的标注语音。本文所述低资源包括了上述定义中的非常低资源和低资源两种情况。部分低资源 ASR 方案的实施需要一定条件,例如文本到语音(Text to Speech,TTS)生成模型能用于增强语音识别系

到稿日期:2024-09-02　返修日期:2024-11-12
基金项目:国家自然科学基金(62266048,62466060)
This work was supported by the National Natural Science Foundation of China(62266048,62466060).
通信作者:杨健(yangjian@yxnu.edu.cn)

统训练数据的多样性，但非常低资源环境下难以构建可用的高质量语音合成系统[7]。低资源可以分为两种类型：1）有标注的语音数据较少，但存在较多（或较容易获得的）无标注语音；2）有标注和无标注的语音数据都较少。这两种情况对应的主流解决途径不同。例如，前者可以结合无标注语音语料和无语音的外来文本数据，采用数据匹配策略和相应的数据增强方法[8-11]；后者可采用基于有标注语音的措施，采取数据增强的方法产生新的标注语音，或用机器学习方法增强现有标注数据的表示能力。

除了标注的语音数据和配对的转录文本较少，低资源语言语音识别经常伴随着方言问题：在关联地域里，多种语言发音和语法相近（甚至是部分融合），但单一语言使用人数少；或者单种低资源语言在不同地域存在部分发音和语法差异。因此，在低资源语言识别中，多语言和方言识别也是重要方向。语音数据的缺乏会导致全神经模型和端到端模型难以充分提取目标语音中的声学特征，可以利用多语言数据预训练的方法提高识别模型的声学特征提取效率。但本文并不专注于模型架构和训练方法的改进。

1.2　相关工作

已有文献对低资源语言 ASR 研究做出总结综述。文献[12]以多语言迁移学习的 ASR 模型为调研目标，研究了低资源环境下多语模型是否优于单语模型、未标注数据的作用、如何构建低资源 ASR 模型等内容，作者认为：交叉语言训练学习到的特征对未出现语言也是可迁移的；迁移学习中如果增加语言 ID 或改进采样，能缓解高资源语言带来的数据不平衡问题；除非多语言模型中存在一种与目标语言发音和语法密切相关的语言，若无，则使用少量有标记的目标语言语料进行模型微调就非常重要。然而，文献[12]主要研究的是低资源 ASR 在多语迁移学习下的模型改进问题，并不是对数据增强的专门调研。文献[13]综述了特征提取和声学模型的研究现状，并对数据扩充进行了调研，但其调研主要面向传统混合模型下的低资源语音识别，而不是面向端到端模型的数据增强。文献[14]探索分析了在低资源环境下一系列 Wav2vec 预训练模型，调研了预训练采用单语言和多语言数据的区别，比较了语音数据利用、多语言学习和音素识别辅助任务 3 个微调方法的作用以及它们在端到端和混合系统中的应用，但其也不是针对端到端低资源 ASR 的数据增强方案的综述。

一些传统语音增强的典型方法，如谱减法（Spectral Subtraction）、基于统计学方法以及 Wiener 滤波等，主要目的是在有噪声的语音环境中提取纯净语音。本文调研的数据增强主要针对“低资源”特征，而不是语音信号的降噪。低资源的基本特征是标注的训练数据不足，语音特征无法覆盖整个语义范围。采用数据扩充或增强方法，是解决低资源语音识别的基本方法。

训练数据不足意味着无法在当前模型下得到足够的语音特征表示。持续的预训练是使语音表示模型适应新语言的最有效方法[15]，预训练和迁移微调的方案被广泛用于多语言和跨语言模型，声学模型共享方案可被用于 0 资源语言语音识别[16]。此外，多任务学习[17]、元学习[18]、元对抗学习[19]等也是低资源语言 ASR 的研究方向。但本文重点不是介绍低资源语言 ASR 模型架构改进，而是聚焦于训练阶段的数据预处理及数据增强措施。本文所指数据增强概念并不局限于狭义的标注数据数量上的增广，而是广义的以提高语音数据表征能力和表征质量为目的的增强。本文将从训练数据的获取和搜集、特征表示和数据重构等不同角度对低资源语言语音识别中数据处理和增强技术进行综述。总体结构如图 1 所示。

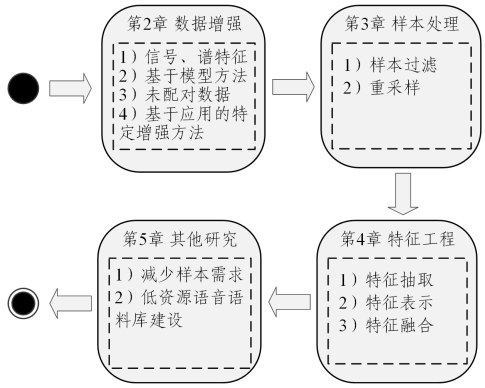


图 1　本文结构

Fig. 1　Structure of this paper

2　数据增强

在低资源语言语音识别中，数据增强主要指数据样本数量增加或表示能力的增强。前者能够提高样本的多样性和鲁棒性，后者则可以在相同数量的数据下使识别模型获得更好的识别性能。数据增强可以粗略分为传统基于信号或谱特征的方法和结合机器学习模型的方法。下面主要从这两方面展开叙述，并对未配对语音和文本的增强以及特定应用环境下的数据增强加以阐述。

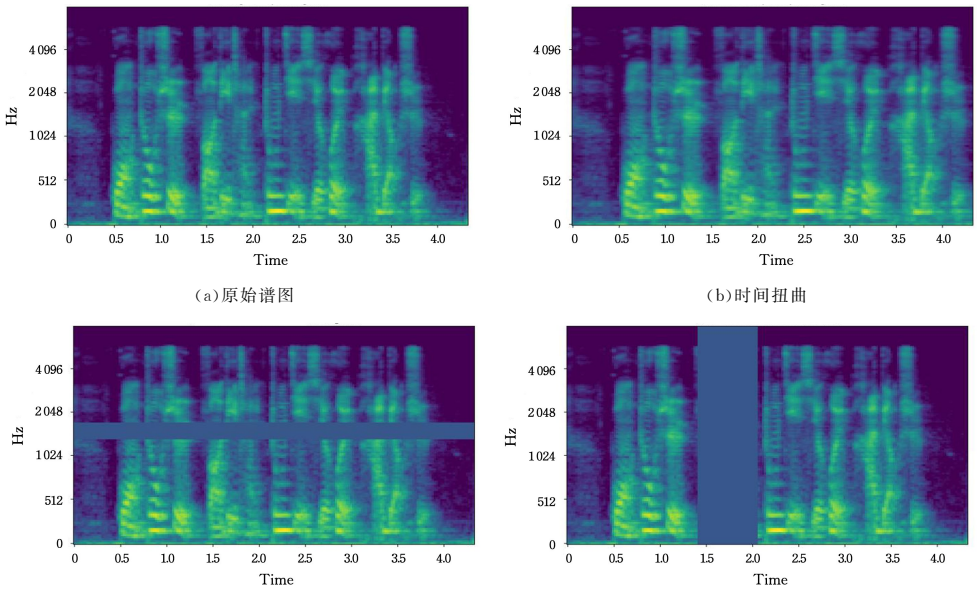
2.1　基于信号或谱特征的数据增强

传统方法主要利用语音的信号特征，如语速、音调、时间和频率维度数据上的扰动扭曲（Warping）、噪声叠加等。原始标注音频经过这些简单变换后，能成倍增加标注训练数据副本。音频速度或音调变化、噪声干扰的简单应用或结合，能有效提高低资源语言识别模型性能[20-21]。文献[22]在使用 TTS 进行数据增强的基础上，叠加噪声，提高了用 TTS 生成数据来训练的模型的泛化能力。文献[23]借助 Python 工具包、房间声学模拟库 Pyroomacoustics 以及 OpenSLR，在原始标注数据上增加白噪声和日常生活噪声以增强数据。

基于谱特征的方法中，最具代表性的、应用最广泛的一类方法是 SpecAugment[24] 及其改进模型[25-28]。SpecAugment 是将音频谱图看作二维（时间和频率）图像，直接对输入特征应用时间和频率上的变换，包括时间维度的特征扭曲、时间和频率维度的块遮蔽等，如图 2 所示。

SpecAugment 在 LibriSpeech 测试集上，相比之前的最好模型，WER 降低了 22.7%[24]。文献[25]对 SpecAugment 中的 3 种策略进行比较研究，发现在使用动态增强策略训练

模型时，频率遮蔽的增强效果总是最大，而时间扭曲最小。尽管 SpecAugment 提出时是在非低资源语音数据库上进行测试的，但 SpecAugment 已被证明在低资源语音识别上也有很好的效果[1,10,25-28]。自 SpecAugment 后，针对不同场景和低资源语言环境，多种 SpecAugment 的改进方案被提出。文献[24]在固定的增强策略中使用 3 种频谱扰动策略对输入语音频谱特征进行增强，固定的增强策略在模型学习阶段不会带来过量的数据多样性，使模型训练可以更快收敛。不同于 SpecAugment 中采用给定频率范围的均匀随机值来替换谱图中的增强区域，文献[27]提出了几个动态遮蔽替换方案：用随机值乘以待增强区域的谱增强(AugMult)；用随机值替换待增强区域数据的谱增强(AugRepl)；连接两个训练样本的原始音频及对应转录文本的输入级联(IC)。与原 SpecAugment 以及其他几个基准增强方法相比，这些数据增强方法配合最新的语音识别模型，在 LibriSpeech-100 数据集仿真低资源语言的条件下训练的 ASR 模型取得了更低的 WER。



(a)原始谱图 (b)时间扭曲

(c)频率遮蔽 (d)时间遮蔽

图 2 SpecAugment 数据增强

Fig. 2 SpecAugment data augmentation

遮蔽时间或频率域的谱信息在低资源 ASR 中会造成数据差异性降低。此外，SpecAugment 中的遮蔽策略是随机的，与模型训练状态无关，从而影响了模型识别质量。针对这些问题，文献[29]提出 SapAugment，根据训练样本的不同损失值来选择在哪些训练样本处进行增强，然后用于下一训练阶段。文献[25]对文献[29]进行改进，提出基于损失函数动态更新策略的 SpecAugment。考虑到模型训练的前一 epoch 损失反映了当前模型状态与前一 epoch 训练采用的增强策略的拟合程度，将损失用于增强策略的选择和增强参数变化因子的计算，目的是鼓励增强方法产生模型训练需要的多样化数据。文献[1]使用了音调和语音速率扰动、SpecAugment 以及三者的组合，得到 3 个修改后的语音数据副本。

数据增强后的语音由于谱泄露可能造成失真，反而降低了 ASR 性能。文献[30]提出带有联合训练框架的两阶段深度谱融合，用于噪声鲁棒的端到端 ASR。SpecAugment 及其改进方法被广泛用于低资源 ASR 的数据增强，但文献[6]中的实验表明，非常低资源条件下使用 SpecAugment 进行数据增强，在性能上并没有明显提升。SpecAugment 及其改进方法是在谱图上进行频率或时间域的遮蔽扭曲，相当于在学习模型上增加一个正则项来提高模型的泛化能力。SpecAugment 的比较、改进方案总结如表 1 所列。

表 1 SpecAugment 的比较以及改进方案

Table 1 Comparison and improvement schemes of SpecAugment

比较与改进方向	文献	内容
比较研究	[6,25,27]	SpecAugment 中 3 种策略增强效果比较，应用背景的比较，与其他增强方法的比较
遮蔽扰动的改进方案	[24,27]	多种谱扰动方案，动态谱扰动和遮蔽方案
启发式的数据增强方案	[25,29]	结合训练(如样本训练损失)动态选择增强方案应用的原始数据
与其他信号增强方案结合使用	[1,26]	结合语速、音调、噪声叠加等方案，这种结合在较多的采用 SpecAugment 增强数据的方法中出现

2.2 基于模型的数据增强

除了利用信号特征，还可以结合机器学习方法来提高语音特征的表示能力。半监督学习、自监督学习都被广泛用于数据的表示学习和增强[4,8,31-34]。还有将多种机器学习方案相结合的方法，例如文献[35]针对低资源环境下攻击性言语检测问题，提出联合对抗学习和迁移学习的协议正则化训练(Agreement Regularized Training)模型。首先采用对抗性训练和 Procrustes 分析，将两个独立语言(低资源和富资源语言)的词嵌入映射到共享空间，在共享空间中基于相似度来生成低资源语言的标注数据。然后设计一种协议正则化迁移学习，使模型从资源丰富的语言迁移到资源贫乏的语言时，仍能保持性能。在基于模型的数据增强方法中，课程学习、语音

转换和建模单元选择是需要详细论述的重要研究方向。

2.2.1　课程学习

在低资源环境下，模型训练中的动态训练数据选择可以提高数据的整体表示能力，其中代表性的是课程学习（Curriculum Learning）[36-37]。根据课程内容的不同，课程学习架构可以分为以数据为中心和以模型为中心两类。因为本文考虑的是数据增强问题，所以主要介绍以数据为中心的课程学习（见图3）。以数据为中心的课程学习不采用随机排列的样本训练，而是按照以难度或复杂性排序的样本顺序（或损失权重）对模型进行训练，提高数据利用效率。其本质是模仿人类学习过程，通过逐渐呈现越来越具有挑战性的训练样本，逐步提高模型的学习能力，更快、更稳定地收敛模型。其核心是训练样本难度的衡量。

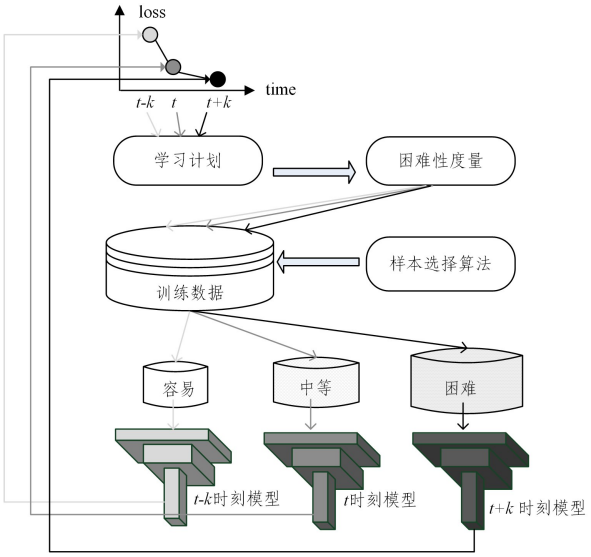


图3　以数据为中心的课程学习

Fig. 3　Data-centric curriculum learning

训练样本排列是课程学习的基础[38]，即采用何种难度度量来决定样本顺序。为了研究样本排列方式对性能的影响，文献[38]探索了多个评分函数，如基于样本持续时间（长度）的评分、基于上轮训练损失的评分、基于ASR性能度量函数（如词错误率WER）的评分，以及后两者的均衡混合评分。这里的均衡混合是指在损失和度量的难度评分基础上，按照一定比例将简单样本与中等和困难样本混合来训练模型。研究结果表明，基于词错误率的均衡混合在几个数据集和模型下都获得了最好的结果。文献[39]为了优化训练样本顺序，基于模型在训练时的进度和训练样例难度的先验知识来优化训练样例的顺序，实现自动化课程学习。该模型引入了一个新的困难度度量——压缩比（Compression Ratio，CR），如式（1）所示。这是对原始音频在不同噪声环境下的评分。考虑到信号噪声越少，信号熵越低，使用gzip等压缩工具的CR值就越大。

$$CR=1-\frac{Size_{\text{after}}}{Size_{\text{before}}} \tag{1}$$

其中，$Size_{\text{before}}$表示音频文件压缩前的大小，$Size_{\text{after}}$表示文件压缩后的大小。

课程学习的学习策略包含学习计划及样本选择方法（见图3），其重点是研究动态的课程学习机制。文献[40]在每个训练阶段后使用一个冻结模型（Frozen model）计算各个样本的得分，并将其作为困难度。在样本得分基础上，还可以使用长度归一化损失（损失除以语料长度）或基于下降速度（前一epoch与当前epoch的得分差距）的比较作为困难度评估，然后依据困难度逐步增加训练样本到以后的训练步骤中。文献[41]提出一种动态课程学习策略作为模型的热身策略和长度扰动策略。考虑到训练中越低的损失输出表示ASR模型能学习得更好，使用冻结模型计算每个训练阶段后所有训练样本的损失，将每个样本的损失作为其难度度量，由易到难逐步增加样本数量。训练数量增加比率如式（2）所示[41]：

$$a(t)=\min\left(1,a_0+\frac{\beta t}{T}(1-a_0)\right) \tag{2}$$

其中，t表示第t个阶段，a_0表示用于训练的数据初始比率，β表示数据增加因子，T表示总的训练次数。

按照式（2）的增量来逐渐增加训练样本，直至覆盖全部训练集，实现渐进式训练策略。动态的课程学习能够充分有效地利用已有的标注数据，是低资源语言ASR研究中的重要方向。

2.2.2　语音转换

在数据增强策略中，一种方法是对某个说话人的语音语料进行处理，使其听起来像是另一个说话人的语料，这也被称为语音转换（Voice Conversion，VC）[42]。语音转换通过增强有限的标注数据来改进低资源语言的语音识别系统，提高系统泛化能力。早期VC技术主要基于平行语料数据集，即多说话人对同一语料发音。但在低资源环境下，难以获得同一语料的多说话人发音，数据增强需要非平行语料的语音转换。深度学习的发展推动非平行语料语音转换从基于语料对齐改进的方法和基于音素后验图的方法，转向基于生成对抗网络[43]、迁移学习[44]等深度学习方法。基于深度学习的编码器能高效地从非平行语音语料中抽取声学特征编码[45]，因而分离说话人相关和不相关特征的基于特征解耦的语音转换成为该领域的研究热点，其适应不同任务背景的特性，非常适合低资源语音识别的数据增强任务。

文献[6]评估了VC系统是否可以跨语言使用来改善低资源语音识别。答案是肯定的，且在富资源语言上训练的VC系统可被用于给从未出现（Unseen）的低资源语言生成额外的训练数据。生成对抗网络可用于实现语音转换和数据增强[1]。在儿童语音识别场景下，可以利用VC从成人语音语料中获得儿童语音标注数据[46]。文献[47]利用一种多阶段学习策略，将语音中语言的音频特征与说话人发音特征相分离，使用前者作为主体特征来训练系统，再添加说话人特征信息以实现说话人自适应的低资源ASR，这也可以看作是一种特殊的VC方法。文献[48]通过对说话人特定变换矩阵和偏置向量的估计来进行原始音频的说话人特征剥离，实现说话人自适应的数据增强。低资源环境下，由于训练数量不足，难以获取足够的不同类型说话人语料，说话人自适应问题是需要着重解决的问题。

2.2.3　建模单元

端到端模型本身对建模单元没有限制，不再受限于混合

框架中声学模型对音素级建模单元的需求，大多数端到端ASR往往采用较粗粒度的建模单元。然而粗粒度建模单元偏向于提供语言学知识，细粒度建模单元倾向于提供更多的声音知识[49]，如图4所示。选择不同的建模单元，会对低资源ASR的性能造成不同程度的影响[50-52]。此外，不同语言特色也影响着建模单元的选择问题。在中文汉语识别中，常用的建模单元有音素、音节、字、子词等。在中国少数民族语言和地方方言识别中，则需要根据具体情况选择合适的建模单元或混合建模单元。例如，在藏语ASR中，可以采用音素或藏文字作为建模单元，也可以建立多粒度的识别模型，不同粒度建模单元的语料可以看作不同语言。将藏文字拆解为更细粒度的部首作为建模单元，一定程度上缓解了语料稀缺的问题[50,53-54]。文献[49]对藏语语音识别建模单元进行比较研究，分析了藏文字的结构和发音特征，在端到端模型的输出词表中包含了藏文字、藏文部首(Radical)、汉字，以及汉语拼音，下游任务可以融合这4种建模单元。

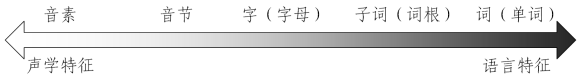


图4 建模单元与特征关联

Fig. 4 Relationship between modeling units and features

实验表明，不同建模单元在识别性能上具有互补性。文献[51]引入了高度压缩、可靠的声学建模子字符单元，结合迁移学习和元学习以提高ASR性能。因为发音属性对于所有人类语言都是通用的，所以也被用于组合的建模单元[55-56]。对于稀缺的印度方言语音识别，文献[52]通过比较原始转录文本、梵文库语音(SLP1)编码和音节来评估建模单元在训练和迁移学习中的表现，使用子词单位，如字符、字节对编码(Byte Pair Encoding，BPE)和一元语言建模(Unigram Language Modeling，ULM)进行标记。实验表明，如果音节多样性在现有语言的音节分布中得到满足，则基于音节的BPE或ULM子词是较好的建模单元。文献[57]分析了维吾尔语的形态结构，认为以词作为识别单元会导致词汇表爆炸，以字为建模单元则会导致序列过长，增加模型收敛的难度，因此采用BPE构造子词作为建模单元。音素BPE编码也被用于低资源的越南语识别[58]。文献[59]提出基于子词的神经文本增强，将生成式大模型的输出文本映射到基于统计特征形成的子词空间中。虽然这种方法并非面向低资源语言ASR，但它显著降低了目标词表大小和存储空间需求。如果这种统计子词空间能够结合聚类算法，与无标注语音和迁移学习方法相结合，在低资源语音识别中也可实现对无标注语音的利用。

词表外(Out of Vocabulary，OOV)词汇的检测恢复也可看作一种建模单元输出的误差修正问题。在NLP中，两级的词加拼写生成语言模型已被证明可以很好地处理涉及OOV的任务[60]。文献[61]将这种架构引入E2E ASR中，使用单词和子词两级建模单元，在LAS(Listen，Attend and Spell)架构中，主要的单词预测网络被训练来预测单词，而次要的拼写网络的优化目标则是基于主要网络的内部表示来预测单词拼写(例如，来自注意力模块的单词嵌入或上下文向量)，这不但可以改进模型的WER，还可以用于OOV词的检测和恢复。对于发音相同但输出词不同的情况，可以采用先约简再重构的方案[62]。首先识别这些语言中声学上相似的字素，使用语义关联的约简来减少ASR系统的中间建模单元规模，然后使用独立的模块重建原始字母。这减轻了低资源端到端ASR系统的学习负担，因为它只需要预测简化字母表。这个工作与ASR的误差修正模型密切相关，因为重建过程相当于一个误差修正。

中文语音识别中，一个不可忽视的问题就是同音字以及书写变体(来源于不同数据集的同一个字的不同写法问题)。文献[63]提出同音外延和统一书写两种方法来解决低资源粤语识别中的这两个问题。前者是在解码搜索中，用一个同音词词典代替具体的词，使得稀有词也能出现在搜索结果中。然后对语言模型重新评分，确定具体的字。后者将广东话的变体合并到最常用的变体中，并增加特定字符的训练样本。这两种方法虽然面向的问题不同，但都是为了解决由于语言特性带来的原始建模单元的复杂性问题(同音或同字不同型)，将建模单元粗粒度化，然后用语言模型加以改进。

2.3 未配对语音和未配对文本的数据增强

未配对的语音和文本数据无法直接用于ASR模型训练，但经过数据处理或是改进特征抽取模型后，其对低资源ASR仍能起到较好的增强作用[64]。对于没有转录的语音，通常方法是构建伪标签(Pseudo Token)，在解码阶段将伪标签与转录文本共同构成搜索空间，再将伪标签映射为真实转录标签。没有语音的文本数据也可用于语音识别的数据增强。主流方案是利用TTS，将文本数据生成对应的语音，以此构建有标注的训练数据。

2.3.1 未配对语音

未配对语音指的是没有标注的语音音频。其用于训练的常见方式是将其声学特征映射为过渡的伪标签，并在后期转换为实际标注。语音特征与地域特征密切相关，使用多个邻近地域的多种低资源语言少量语音数据的组合语料来训练多语言识别模型，能最大限度地降低数据稀缺的负面影响。文献[65]利用多印度语字符集间发音相似性，使用通用的梵文图书馆语音编码作为过渡标签，构建几个南部印度语的单一ASR模型。这个过渡标签集通过简单查找表和预定义规则就能映射为目标语言的转录。除了使用多语言通用以及组合的编码集以外，伪标签也可以在声学模型建模过程中构建，如使用注意力特征[66]。文献[67]利用Transformer输出和源注意力权重来构造包含每个Transformer输出的后验(概率)和时序信息的伪目标标识，并借助这个伪标识构造知识蒸馏模型(Knowledge Distillation，KD)，以提高CTC-Transformer网络中共享编码器的特征提取能力。KD被广泛用于语音应用中的模型压缩和域适应[68]，它构建一个轻量化的小模型(学生模型)，利用训练更完备的大模型(教师模型)的监督信息来训练这个小模型(见图5)，以获得更好的性能和精度。

在语音情绪识别任务中，自监督学习(Self-Supervised Learning，SSL)被用于在新语言未标注话语上生成伪标签[34]。采用SSL输出的概率分布向量作为硬伪标签(Hard Pseudo Label)，在其上增加两个正则项后作为软(Soft)伪标签，研究了两种伪标签在低资源ASR下的性能。结果显示，

前者综合性能更好，这可能是由于混合数据增强给数据带来过多的噪声，再应用正则项反而会降低性能。另一方面，也说明低资源环境下 ASR 系统的泛化能力需要谨慎对待，应采用适当的正则化措施。

除了伪标签以外，音素建模单元也可以作为过渡标注单元。在面向低资源 ASR 的迁移学习模型中，低资源的目标语言可能出现未见的语料。从已有语言到未出现语言的 ASR 应用中，一个关键步骤就是创建未出现语言的音素表，以音素作为过渡建模单元。文献[69]以无监督方法在训练中创建音素词表，不需要任何预备的语言知识（调研、分析、方法）。对于只有口头发音，没有文字的低资源语言，音素作为过渡识别标签的方案有着重要的意义：大量语言学研究中，对于这类语言的调研或档案文献，通常都是使用国际音标（International Phonetic Alphabet，IPA）记录发音，使用通用的富资源语言记录语法含义。这类语言无法构建语音识别应用，但利用音素作为中间标注，可以构建低资源语言发音到高资源语言转录的翻译或互译系统，有助于低资源语言的保护、传承和文化交流。

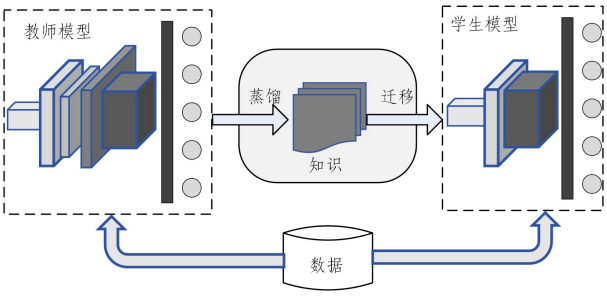


图 5　知识蒸馏的基本架构

Fig. 5　Basic framework of knowledge distillation

2.3.2　未配对文本

未配对文本能提高跨语言迁移学习能力[70]。训练 ASR 模型可以与一组文本到文本的辅助任务（如机器翻译）联合起来[71]，共享一个解码器和部分编码器，以此提升 ASR 训练效果。迁移学习中，基于富资源语言预训练的初始 ASR 先建立在相对较小的领域中，然后逐渐增加领域内低资源目标语言训练数据，来改进性能。这个过程被称为自举（Bootstrapping）。文献[10]提出一个文本到文本的映射方案，以实现轻量级低成本的 bootstrapping，使用少量真实音频与 TTS 音频混合训练，以提高自举 ASR 性能。文献[72]提出另一种利用未配对文本的思路。首先利用标注语音生成语音特征与对应文本字形单元的对照字典。然后参考未配对文本，从对照字典中查询匹配的语音特征并拼接，构成合成的标注语料。未配对文本可用于预训练模型，再借助跨模态知识传输学习框架[73]，使用带有语言学表达的层次化声学对齐方案，将语言知识转换为声学编码用于 ASR。

低资源环境下，对于未配对文本，通常利用 TTS 模型（如 WaveGAN，SpecGAN，Tacotron 2）将其合成音频。实践表明，TTS 合成音频数据会给模型带来更好的离散度（区分度），避免模型过拟合[9-10,74]。文献[9]在基于 Transformer 的 ASR 识别结果的纠正任务中，利用文本到语音的合成数据来提高训练数据的多样化。文献[75]使用文本到语音系统生成 OOV 词，并调整损失以鼓励神经网络更多地关注 OOV 词。文献[28]收集外部文本数据进行语言建模，并训练 TTS 模型来生成语音文本配对数据；基于传统的混合结构构建系统，并使用不同的声学神经网络架构和不同的数据增强方法开发了各种子系统；最后，采用系统融合得到用于训练的增强数据。文献[47]提出一种多阶段学习策略，其目标是提高对未见说话者的 0 命中话语自适应能力。该方案利用基于部分网络的深度迁移学习，在将高资源语言作为源域的基础上使用预训练的单语单说话人 TTS 和 d-vector 说话者特征编码器[76]来克服低资源问题。d-vector 是一个固定长度的嵌入向量，用于表示说话人特征。文献[47]提出的方法的本质是将语音中的音频特征与说话人个人发音特征分离，使用前者训练模型以提高低资源 ASR 系统的声学特征抽取能力，使用后者提高模型的说话人自适应能力。文献[77]利用一个面向高资源枢轴语言（Pivot）训练的 TTS 系统，依据目标语言文本生成合成语音。作者调查了这种技术何时以及如何在低资源环境中最有效。实验结果表明，随着 TTS 合成音频训练数据的增多，低资源 ASR 的性能增强，当合成音频与真实枢轴语言音频数据量相当时，达到最好性能。但 TTS 数量到达某个拐点后，性能反而会下降。此外，合成音频质量对迁移后低资源 ASR 的识别，并没有较强的正向促进作用。

语音翻译是将源语言语音翻译成目标语言文本的过程。低资源语音翻译现有研究大多通过 TTS 增强源语言语音，只有少数研究者关注目标语言文本增强。文献[78]提出一个目标端数据增强方法。端到端语音翻译的任务是输出目标语言的语义表达，由于缺乏配对语料，因此可采用意译（Paraphrase）结果作为目标端输出。意译作为自然语言中的一种常见现象，类似于文本摘要，与语言变异性的核心内容密切相关。文献[78]提出通过意译目标端句子来扩展端到端语音翻译的训练数据。首先基于一个释义生成模型，结合几个统计机器翻译特征和通常使用的 RNN 特征生成目标端意译句子，然后利用一种结合文本语义相似度和语音-文本协同评分的过滤模型，过滤未对齐的语音-文本对，最后将语音与匹配的释义文本构成低资源语音翻译模型的训练数据。

2.3.3　同时使用未配对语音和文本

未配对语言和未配对文本能够同时用于低资源语言 ASR 模型训练。文献[8]提出一种联合交替训练方法，同时利用未配对语音和文本来训练一个通用 ASR 模型。前者通过一个经过预训练和微调的 ASR 模型来生成语音-伪标签配对数据集，后者通过一个 TTS 来合成音频进而构造伪语音-标签集合，从预测概率分布和语音特征提取两个方面进行分析，说明了二者在语言标记预测和声学特征学习方面是互补的。训练时，使用伪标签的遮蔽解决不正确标签的过拟合问题，构造音频特征敏感的并行层缓解合成音频特征提取的不准确问题，从而提高交替训练的效率。

2.4　特定应用背景下的数据增强

在特定应用领域中，未配对语音和文本也可用于提高语音识别性能，如利用同声翻译数据库。在国际会议和法庭诉讼等多语种活动中，资源贫乏语言的发言通常可由人进行同声传译。在这种情况下，可以假设同声翻译的内容与原语音

的转录内容相同。在这种场景下，文献[79]构造了一个 ASR 和机器翻译的同步过程，并使用 ASR 编码器和机器翻译编码器的交叉注意力机制的联合架构来实现低资源语言 ASR。文献[47]基于儿童语音和成人语音的不同声学特点，利用声道长度扰动(Vocal Tract Length Perturbation，VTLP)[80]和基于信号处理的数据增强方法，在成人语音上应用特定滤波函数来构造儿童语音语料，进行数据增强。文献[48]针对口腔疾病患者的语音识别开展研究，原始数据从 YouTube 上搜集，得出某一类音素比其他音素更容易被 ASR 识别，在下游应用中可以利用这一点，例如构建语音助手时可以选择主要由这一类音素组成的语音命令。此外，在声学建模时，通过对说话人特定变换矩阵和偏置向量的估计来进行说话人特征剥离，实现说话人自适应的数据增强。

2.5 小结

传统基于信号和谱特征的数据增强通过修改信号本身来获得多样化的训练样本，以缓解低资源环境下的训练样本数量问题。基于模型的方法则利用机器学习方法进一步提炼语音和转录数据中的知识，或将知识抽取模型与 ASR 模型有机结合，提高低资源语言 ASR 的性能。对于一般结构 ASR 系统来说，未配对语音和样本并不能用于模型训练，但通过伪标注、迁移学习及转录文本融合、TTS 系统辅助生成标注音频训练样本等方法，可以将其有效应用于低资源语言的 ASR 系统训练。这些方法还能综合起来进一步提高系统性能。低资源语言的数据增强可以用知识图谱来表示，如图 6 所示。

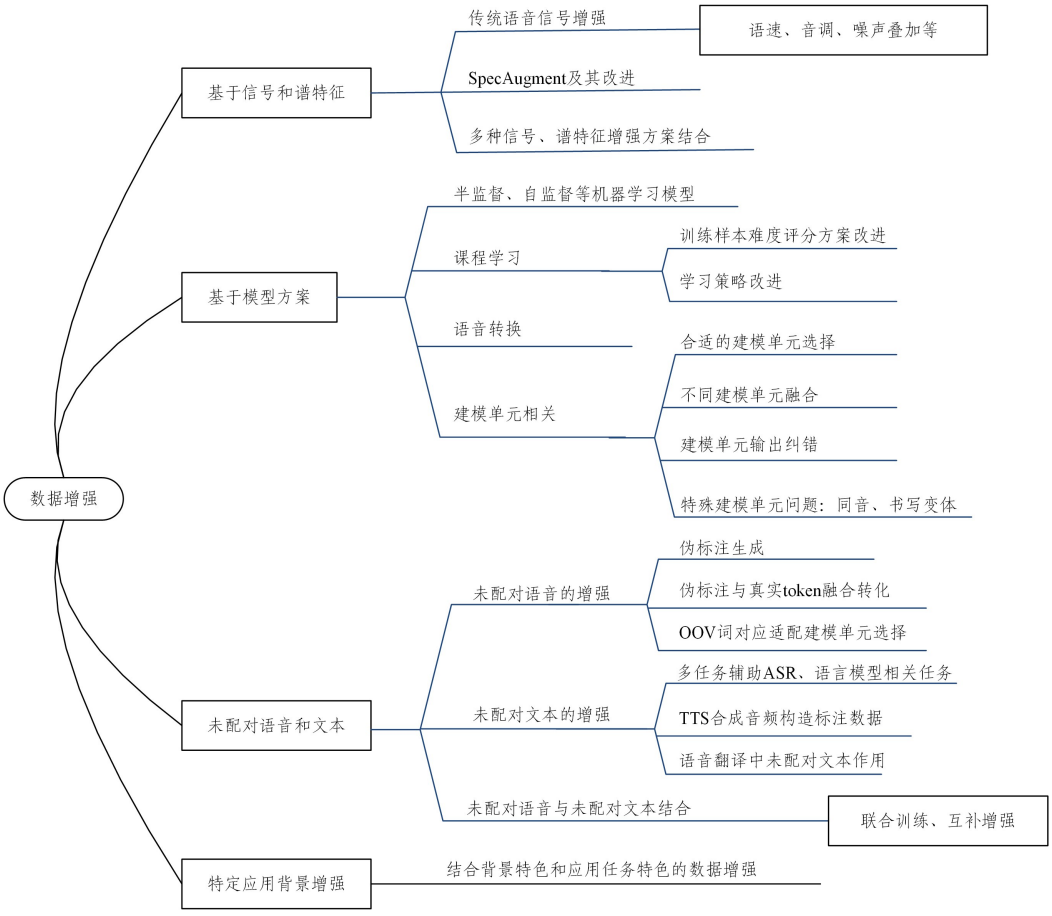


图 6 数据增强方案的知识图谱

Fig. 6 Knowledge mapping of data augmentation schemes

3 样本过滤和重采样

原始的低资源语音音频信号需要经过处理，才能具有较好的特征表述能力，其中对比学习是一个热点研究方向。例如，对比预测编码(CPC)是从高维序列中抽取有用标识信息，利用一个编码器把数据中抽取的上下文表示与未来样本表示进行对比，使得编码器能够学会预测未来的最佳表示。在语音识别中，CPC 通过学习对两个样本语音帧的相似(正样本)和不相似(负样本)的编码来构造语音帧的特征表示[81-82]。对比学习中需要构造与正样本对应的负样本，负样本的采集过滤质量决定了编码模型的质量。针对负样本集过滤问题，文献[31]提出 FNIE 方法来过滤假阴性样本，提高语音中阴性样本集的质量。FNIE 更支持向量与负样本向量，并优化对比学习中常用的 InfoNCE 损失函数，允许模型学习更好的语音表达。样本过滤的增强方法还包括在原始信号上应用特殊的滤波函数。例如，在儿童语音识别场景中，由于缺乏儿童标注语音数据，系统性能通常较差。声道长度扰动(VTLP)是弥合成人和儿童说话者语音差距的最常见方法[80]，它是通过沿频率轴缩放特征向量来实现的，模拟了不同声道长度对语音信号频谱包络的影响。为了缓解儿童语音数据不足的情况，文献[46]引入了一种数据增强策略。该策略对输入信号的源和滤波器分量应用特殊的翘曲函数，将成人语音谱图沿频率轴应用这个函数，使语音信号反映的声道长度特征更接近儿童，从而利用成人语音实现儿童语音数据增强。在多语

言预训练中，文献[40]首先构造一个语言分类器，然后使用语言分类器中提取的目标语言后验对训练样本进行数据加权，使得模型在预训练过程中更偏向于目标语言，从而提高 ASR 模型在目标语言上的识别精度。

低资源语言样本数量少导致复杂模型容易过拟合，下采样是一种可行的解决方案[67,83-86]。文献[87]针对低资源语言语音识别中语流音变(Pronunciation Variation)造成的影响，基于语言学家在发音变化上的研究成果，构建了发音差异处理模块，直接从语音谱图上抽取特征信息以捕获发音变化特征，并通过多头注意力机制和改进的 GRU(Gate Recurrent Unit)模块有效融合纯语音和谱特征，突出共振峰差异，缓解数据稀缺造成的模型过拟合风险。下采样会造成信号失真，降低语音信号表达能力，采用膨胀卷积(Dilated Convolution)能够在保留长距离依赖的条件下获得更短的语音信号表达[88-90]，更有助于在低资源条件下提升 ASR 精度。在多语言预训练和迁移学习基础上，面对新出现的低资源语言，文献[34]随机重复采样新语言标注语料，缓解语料数量不平衡问题；还利用模型改进下采样方法，例如在传统的下采样方法的基础上，在 Transformer 编码器层内部加入时间缩减层，进一步降低帧率[86]。文献[41]根据单词边界对语料进行切片，并创建增强样本。其本质就是在考虑对齐的情况下，将一个长语料切分成若干短的样本，从而增加样本数量。

4　特征抽取、嵌入、融合

语音信号包含丰富信息，如音素、韵律、语种、语音内容、说话人特征及情感等，而语音识别中常用的特征包括梅尔频率倒谱系数(MFCC)、感知线性预测系数(PLP)、Fbank(滤波器组)、Spectrogram(频谱图)等。低资源语言 ASR 要尽可能利用语音信号中的有用信息，使用传统 ASR 的常用特征是不够的，需要相应的改进或进行特征融合以适应低资源环境。

4.1　特征抽取

低资源环境的语音特征如表 2 所列。语流音变反映了相邻音素的相互影响作用，其差异性主要体现在元音和复合元音共振带的分布上。由于没有足够的标注数据来支撑音素框架对齐，因此传统低资源 ASR 难以捕获语流音变特征。针对这个问题，文献[87]提出从语谱图上抽取特征信息以捕获发音变化特征，并进一步通过多头注意力机制和改进的 GRU 模块有效融合纯语音和频谱特征，突出共振峰差异，缓解数据稀缺造成的模型过拟合风险。

基于狄利克雷过程高斯混合模型(Dirichlet Process Gaussian Mixture Model，DPGMM)[91]的非参数聚类通常应用于口语术语检测和零资源任务等任务，但在大词汇量连续语音识别中尚未得到应用[92]。文献[92]比较了各种特征的可判别性，其中 DPGMM 聚类的语音后验概率具有较强的可判别性。他们通过附加 DPGMM 后验概率来提高声学特征的可分辨性。考虑到婴儿能够通过听无文本的言语而做出适应性的感知，导致永久性的大脑状态改变以形成终身语音感知的物理基础，DPGMM 和 DPGMM-RNN 混合模型被用于提取这种感知特征[93]，以增强 ASR 的声学特征提取能力。文献[94]利用潜回归贝叶斯网络(Latent Regression Bayesian Network，LRBN)进行特征抽取：传统特征被用作 LRBN 的输入，LRBN 的潜层代表学习到的语音表征，其参数学习采用硬期望最大化(Hard Expectation-Maximization，HEM)算法进行训练。不同于计算昂贵的大型模型(如 Wav2vec 2.0)，LRBN 是一种用于学习数据分布和高层特征的轻量级无监督学习模型。

表 2　低资源环境的语音特征

Table 2　Speech features in low-resource scenarios

特征或抽取方法	作用与特点
语流音变	反映相邻音素的相互作用和影响
DPGMM	提高声学特征分辨性
LRBN 潜层特征	轻量级高层声学特征表达

4.2　嵌入表示

隐藏单元聚类框架可用于从原始音频中进行自监督表示学习[95]。输入由带窗口的音频样本组成，并使用一维卷积层进行处理，处理后的“时频”表示再经过长短时记忆(LSTM)层，从而为每个窗口段生成上下文嵌入向量。文献[89]在低资源场景下语言未定的预训练声学模型进行口语意图理解的任务中，对 3 种不同的嵌入(单元)，即音素嵌入、泛音素(Panphone)、Allo 嵌入进行比较研究。首先使用预训练的通用音素解码器 Allosaurus[96]，从原始音频向量中提取出音素解码向量及 Allosaurus 的最后一层输出向量。这两个结果向量再经由一个全连接网络层，构造出上述 3 种不同嵌入单元。其中音素嵌入是将每个音素解码向量映射为固定 256 维向量，泛音素则映射为 26 维向量(对应 26 个英文字母)，而 Allo 嵌入则直接使用 Allosaurus 的最后一层输出作为分类模型嵌入。实验表明，Allo 嵌入具有与语言无关的特性，结合 1-D 膨胀卷积，在低资源语言意图分类任务上获得了 SOTA 结果。文献[53]研究了两种不同的说话人信息嵌入：利用基于 TDNN(Time Delay Neural Network)的预训练网络抽取 512 维的说话人信息嵌入向量 x-vector，以及基于 Transformer 架构的 s-vector。然后在基于 Transformer 的模型中，加入说话人嵌入向量，实现说话人自适应的 ASR。在低资源语言 ASR 中，由于数据缺乏，获取不同说话人语料来训练说话人嵌入向量以实现说话人自适应，反而成为一个困难问题。一种思路是去除原始信号中说话人的特征信息，只保留语音到语义的特征。

Wav2vec 是一个常用的基于自监督学习的嵌入向量获取模型。文献[4]比较了两种利用 Wav2vec 2.0 的方法：基于注意力的端到端模型(AED)以及混合隐马尔可夫模型(HMM/DNN)语音识别系统。前者的 Wav2vec 2.0 用于构造模型编码器，后者的 Wav2vec 2.0 被用作声学建模。实验结果显示，在低资源场景下，利用 Wav2vec 2.0 嵌入的 HMM/DNN 模型优于 AED 模型，但 AED 模型能有效提取全局注意力，在未见说话人语音上更加鲁棒。

4.3　特征融合

特征融合是将不同特征按一定方式组合起来作为 ASR 或下游任务模型训练的输入。按照融合架构的不同可以分为垂直融合和水平融合两种类型。垂直融合是对原始语音信号在一个模型不同位置产生的细粒度和粗粒度或全局与局部特

征的融合;水平融合是原始语音信号通过不同通道获得多个特征表示,再将这些特征以类似拼接或加权和的方式组合。多语言 ASR 中,语言类别特征和领域类别特征可以以水平[97-98]或垂直[99-100]的方式融合入模型的编码器和解码器。文献[97]首先用 Transformer 架构训练语言分类器以获得语言类别嵌入向量,然后将嵌入向量添加到对应语料输入特征中训练多语言 ASR。文献[99]提出一个多尺度语音情感识别方案。识别模型的输入融合了细粒度帧级特征与粗粒度语料级深层特征,同时,基于连接注意力机制(AMSNet)为不同类型特征分配不同权重,充分利用不同类型特征的优势提高模型情感识别的综合性能。文献[100]提出一个利用全局和局部信息相结合的多尺度多通道特征提取和融合结构,获得一个全面综合的语音特征。语音信号的 MFCC 和 ZCR 特征被输入特征抽取模块,特征抽取模块使用 3 个不同通道分别提取信号的浅(Shallow)特征、中度(Moderate)特征和深度(Deep)特征,输出的 3 种特征拼接后,再采用一维卷积神经网络进行特征学习和情感识别,从而获得更好的学习能力和改进的语音情感识别结果。文献[101]提出了一个 BERT 风格的语言模型,称为 Phoneme-BERT。它结合语言模型、音素序列及 ASR 转录文本以学习音素感知的表达。预先训练的 Phoneme-BERT 可以在缺乏音素序列的低资源下游任务中作为纯的单词编码器,并且产生比纯单词的语言模型更好的结果。从融合对象来看,水平融合涉及的被融合特征包括音频的语言类别、方言类别及领域信息,可以看作是引入音频信号的外部辅助特征;垂直融合则通过模型或算法提取音频本身不同粒度特征或模型的不同层次特征,加以融合后参与下层任务。从融合方式来说,有简单的拼接和加权拼接,还有将不同特征经过嵌入后再进行加权线性和,从而获得多特征的融合向量用于识别任务[98]。总而言之,尽可能利用有限语料中的有用信息,是低资源 ASR 需要解决的一个根本性问题。

低资源场景经常会利用其他富资源语言构建多语或交叉语言迁移学习框架,这些框架广泛采用自监督学习来获得语音的声学特征表示,但自监督学习在低资源环境下存在数据和语言不匹配的问题[102]。为了解决这个问题,考虑到谱特征与领域和语言无关,并且它们在多语言模型中的使用表明它们能够实现强大的跨语言迁移,文献[84]构建了一个可学习和可解释的框架来结合谱特征与自监督学习模型的表示输出,提出线性、卷积和基于共同关注的 3 种融合方法。提出的框架在 3 个低资源数据集上的 ASR 和语音翻译任务方面明显优于基线模型和只有自监督学习特征的模型。

5 其他工作

5.1 减少模型的数据需求

基于神经网络的深度模型训练要求海量的训练数据,在低资源环境下,可以考虑在识别性能允许的情况下,改变模型架构和语音信号的嵌入表达,降低识别模型对数据量的需求。文献[103]基于信息不确定性的数据选择,减少标注消耗,其本质是减少模型对标注数据的需求。初始语料集被分为 3∶7 的训练集 $D1$ 和增强集 $D2$ 两个集合,$D1$ 被用于微调一个预训练模型,然后,预训练模型在 $D2$ 上执行认知不确定性采样(Epistemic Uncertainty Sampling)。根据不确定性词错误率,将 $D2$ 中 k 个最不确定的样本移至 $D1$,再用新的 $D1$ 进行下一轮微调。文献[104]中针对低资源语言,提出了一种多语言环境下 DNN-HMM 混合声学模型融合的新方法。针对目标语言语音信号,将不同单语声学模型的后验分布融合在一起。对每个源-目标语言对单独训练一个回归神经网络,将源声学模型的后验信息转换为目标语言。与其他基于神经模型的端到端 ASR 相比,这些网络需要的训练数据量大大减少。

在低资源 ASR 下,KD 可以用于降低模型对标注语音数据的需求量,提高少量标注语音的声学特征表达能力,因而被广泛用于低资源语音识别的数据增强上[67,86,104-106]。在残疾语音识别问题上,文献[105]提出一种阶段性 KD 方案,从教师模型中学习语音信号映射到语义表示的关键特征,避免模型对语音的清晰度和风格等辅助特征过拟合。首先用多个困难和健康说话者的混合语料来训练学生模型,然后用目标困难说话者的训练集进行调优。文献[106]提出一个基于注意力的多层特征蒸馏以自动学习从所有教学模块中总结的特征表达。

5.2 低资源语言语音语料库建设

解决低资源语言语音识别问题的另一途径是扩充规范录音和标注的真实语音语料。因此,探索高效的、节约成本的语料库构建方法,是非常关键且必要的。标注数据较少的原因有很多,如经济文化等客观因素导致语音数据获取困难,没有规范的、公开的语料数据库或数据库建设标准,语言本身没有相应标记符号(只能口口相传或用 IPA 记录语音),语音数据标注成本过高等。针对这些问题,如何低成本、高效建设低资源语音语料库成为一个研究方向。可以采用众包方式,充分利用网络联结分散的说话人和语言、标注专家资源,以松耦合方式逐步建设低资源语音语料库。文献[107]详细阐述了以众包方式编制低资源语言语音数据库的整个过程,从主题设定、语料整理,到利用众包软件进行录音和信息采集,最后进行语音过滤形成数据库。文献[51]阐述了国内开展拉萨方言语音数据库建设的工作,引入了 NICT-Tib1 数据库,这是一个新的开源数据库。研究者分别在单语言和多语言设置下进一步地更新基准系统。文献[108]也论述了白语语音语料库建设方案和实施过程。

除了利用移动 App 或网络应用程序减少语料录音和人工标注成本,还可以充分利用自动标注技术减少标注成本,例如深度主动学习(Deep Active Learning)[109]结合了深度模型提取特征的能力和主动学习能选择信息熵最大的样本构建训练集的能力以及优秀的标注样本的能力。通用音素解码器 Allosaurus[96]可以以黑盒模式将输入的原始音频信号解码为音素序列,这可以作为自动标注工具的组成。还有通用的数据标注工具如 Universal Data Tool[110]可以实现音频和文本数据的自动化/半自动标注。得益于字音转换(Grapheme-to-Phoneme,G2P)以及文本到 IPA 转写方法的成果,如 Phonetisarus[111]、Epitran[112]和许多语言的开源 G2P 模型[113],手工创建词汇的问题得到极大程度的缓解。中文自动标注工具,如 Jieba,THULAC 等,主要集中在提供分词和词性标注上,但从中文语音到音素或音节的自动标注工具较少,可以探索

使用通用音素解码器(如 Allosaurus)在中文或汉藏语系语言上的应用和改进。

目前不存在通用的、面向语音识别及相关任务的语音语料库构建框架。各语音库采用自己的软件架构、存储结构和使用规范,使得低资源语言语音和文本语料库构建成本过高。因此,通用语料库架构和研究及其应用模式的探索是必要的。本文提出一个通用的语音语料库建设框架,如图 7 所示。框架包含软硬件(工具)建设、语料搜集方法与环境、增量式语料管理、分布式的专家/人工标注和自动标注架构、语料审核确认、可持续的语料研究与应用模式等。通用语料库的形成,有助于低资源语言以低成本模式构建并搜集、标注数据,从而推进低资源 ASR 的研究与应用进程。

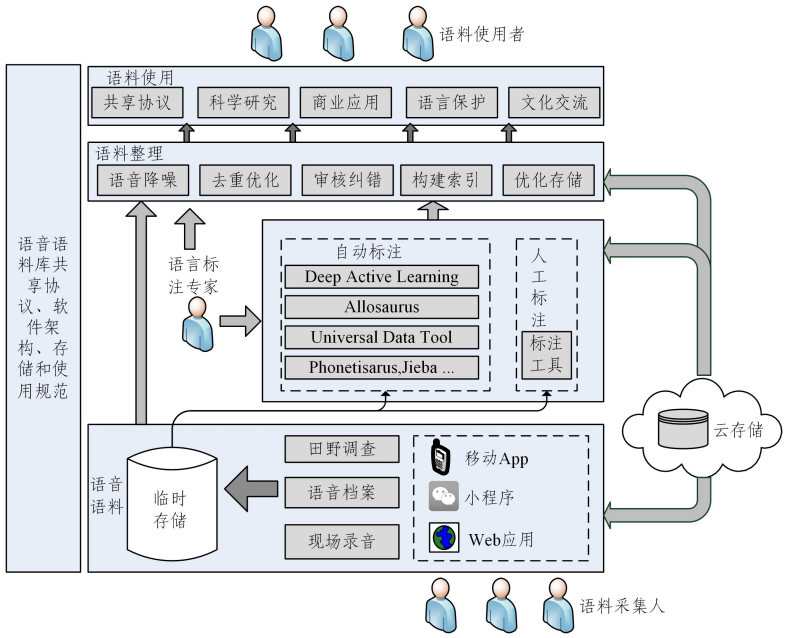


图 7　低资源语音语料库建设框架

Fig. 7　Construction framework of low-resource speech corpus

结束语　随着深度学习技术的发展,语音识别研究成果已进入商业应用,但低资源语言的语音识别由于训练数据量不足的问题,仍然难以在实际中应用。针对这个问题,本文在低资源语言自动语音识别的背景下,从数据增强、样本过滤采样,到模型训练时的特征工程等方面对现有研究成果进行调研和总结。从调研结果来看,低资源语言离大范围成功应用的目标还较远。在方言和语音变体较多的多民族聚集区域,由于社会经济文化的交融和影响,大量低资源和极低资源的语言面临消失的危险,因此需要加快低资源语言数字化进程,探索通用的商业化的面向低资源和极低资源语言的语音搜集、标注、整理和识别技术。通过对现有研究的总结,本文认为未来的研究可以聚焦于以下几点:

1)从数据增强角度。一方面,探索标注数据增广策略,对已有标注数据和多种增强方法进行融合和详细的比较研究,探索适合不同数据量和不同类型语言的优秀方案;另一方面,研究音频声学特征的高效抽取和利用,以及数据特征抽取、多特征融合与模型复杂度的匹配关系,寻找启发式的特征融合策略。此外,无标注语音和无标注文本在低资源 ASR 中的充分结合和利用,是一个需要重点关注的方向。

2)语音增强的各种方法中,缺少针对低资源环境下噪声的作用评估,使用信号扰动的方法对音频信号进行处理,可能会导致信息丢失。在这种情况下,噪声环境的 ASR 系统性能会急剧下降。环境噪声叠加对低资源语言 ASR 系统的数据增强作用和性能影响还需进行进一步探索。研究低资源环境下噪声鲁棒的数据增强方法和相关模型,是一个重要的方向,例如探讨文献[30]的两阶段谱融合算法与低资源数据增强方法的结合。

3)探索低资源语言语音语料库的构建规范,通用语料库软硬件开发,语料共享和利用机制。低资源语言 ASR 的根本问题还是数据量和标注问题,因此,充分利用网络语音搜集、分布式众包的人工标注以及自动化的标注、语音和标注语料的自动化处理,是低资源语音识别可以考虑的应用研究方向。

4)构建针对下游特定任务的特有模型。考虑人类学习语言的过程,从独立的词开始,往简单的语义模型递进,在有教师指导的情况下,进一步探索更多的词和复杂语义模型,以及口音的改进。因此,低资源 ASR 系统利用课程学习,不能仅从数据和模型复杂度构建课程,还应探索如何用课程来体现渐进式的语义表达。

5)除了一些特有应用场景外,现有低资源 ASR 研究较少考虑说话人特征问题。未来应探索说话人嵌入信息在低资源环境下与其他方案融合,以及说话人自适应的低资源 ASR 模型研究。

参 考 文 献

[1] KIPYATKOVA I,KAGIROV I. Deep Models for Low-Resourced Speech Recognition:Livvi-Karelian Case[J]. Mathematics,2023,11(18):1-21.

[2] JOSHI P,SANTY S,BUDHIRAJA A,et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World[C]// Proceedings of the 58th Annual Meeting of the Association for

Computational Linguistics(ACL'20). 2020:6282-6293.
[3] SEBASTIAN R. The 4 Biggest Open Problems in NLP [EB/OL]. (2019-01-18) [2023-12-15]. https://ruder.io/4-biggest-open-problems-in-nlp/.
[4] ROUHE A, VIRKKUNEN A, LEINONEN J, et al. Low Resource Comparison of Attention-based and Hybrid ASR Exploiting wav2vec 2.0[C]// Proceedings of the 23rd Annual Conference of the International Speech Communication Association, Interspeech 2022. 2022:3543-3547.
[5] HEDDERICH M A, LANGE L, ADEL H, et al. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021: 2545-2568.
[6] BAAS M, KAMPER H. Voice Conversion Can Improve ASR in Very Low-Resource Settings[C]// Proceedings of the 23rd Annual Conference of the International Speech Communication Association. ISCA, 2022:3513-3517.
[7] ZHANG J L, MAIRIDAN W, GULANBAIER T. Review of Speech Synthesis Methods Under Low-Resource Condition[J]. Computer Engineering and Applications, 2023, 59(15):1-16.
[8] DU Y Q, ZHANG J, FANG X, et al. A Semi-Supervised Complementary Joint Training Approach for Low-Resource Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31:3908-3921.
[9] LI W, DI H, WANG L, et al. Boost Transformer with BERT and Copying Mechanism for ASR Error Correction[C]// Proceedings of International Joint Conference on Neural Networks. IEEE, 2021:1-6.
[10] GIOLLO M, GUNCELER D, LIU Y L, et al. Bootstrap an End-to-end ASR System by Multilingual Training, Transfer Learning, Text-to-text Mapping and Synthetic Audio[C]// Proceedings of 22nd Annual Conference of the International Speech Communication Association. 2020:2416-2420.
[11] JIN H. Multimodal Enhancement Techniques for Low-Resource Cross-Language Speech Translation Scenarios[D]. Foshan: Foshan University, 2022.
[12] HEMANT Y, SUNAYANA S. A Survey of Multilingual Models for Automatic Speech Recognition[C]// Proceedings of the 13th Language Resources and Evaluation Conference. Marseille France: ELRA, 2022:5071-5079.
[13] SLAM W, LI Y N, UROUVAS N. Frontier Research on Low-Resource Speech Recognition Technology[J]. Sensors, 2023, 23(22):1-47.
[14] ZHAO J, ZHANG W Q. Improving Automatic Speech Recognition Performance for Low-Resource Languages With Self-Supervised Models[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6):1227-1241.
[15] KAROL N, MICHAL P, KYOKO M, et al. Adapting Multilingual Speech Representation Model for a New, Underresourced Language Through Multilingual Fine-tuning and Continued Pretraining[J]. Information Processing & Management, 2023, 60(2):1-12.
[16] WANG H Y, JEON E, ZHANG W Q, et al. Zero Resource Korean ASR Based on Acoustic Model Sharing[J]. Journal of Data Acquisition and Processing. 2023, 38(1):93-100.
[17] D'SA A G, ILLINA I, FOHR D, et al. Exploration ofMulti-corpus Learning for Hate Speech Classification in Low Resource Scenarios[C]// Proceedings of 2022 International Conference on Text, Speech, and Dialogue. Springer, 2022:238-250.
[18] SINGH S, WANG R, HOU F. Improved Meta Learning for Low Resource Speech Recognition[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE Computer Society, 2022:4798-4802.
[19] CHEN Y Q, YANG X K, CHEN Q C. Meta Adversarial Learning Improves Low-Resource Speech Recognition[J]. Computer speech & language, 2024, 84:1-12.
[20] XU F, DAN Y J, YAN K Y, et al. Low-Resource Language Discrimination toward Chinese Dialects with Transfer Learning and Data Augmentation[J]. Transactions on Asian and Low-Resource Language Information Processing, 2021, 21(2):1-21.
[21] KO T, PEDDINTI V, POVEY D, et al. Audio Augmentation for Speech Recognition[C]// Proceedings of INTERSPEECH 2015. 2015:3586-3589.
[22] FAZEL A, YANG W, LIU Y L, et al. SynthASR: Unlocking Synthetic Data for Speech Recognition[C]// Proceedings of INTERSPEECH 2021. 2021:896-900.
[23] YU K. Research on Low-resource Mandarin Dialect Speech Recognition Method and Application[D]. Xi'an: Chang'an University, 2021.
[24] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition[C]// Proceedings of INTERSPEECH 2019. 2019: 2613-2617.
[25] LI R, MA G, ZHAO D, et al. A Policy-based Approach to the SpecAugment Method for Low Resource E2E ASR[C]// Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. 2022:630-635.
[26] ZEYER A, BAHAR P, IRIE K, et al. A Comparison of Transformer and LSTM Encoder Decoder Models for ASR[C]// Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. IEEE, 2019:8-15.
[27] DAMANIA R, HOMAN C, PRUD H E. Combining Simple but Novel Data Augmentation Methods for Improving Conformer ASR[C]// Proceedings of INTERSPEECH 2022. 2022: 4890-4894.
[28] ZHONG G, SONG H, WANG R, et al. External Text Based Data Augmentation for Low-Resource Speech Recognition in the Constrained Condition of OpenASR21 Challenge[C]// Proceedings of INTERSPEECH 2022. 2022:4860-4864.
[29] HU T Y, ASHISH S, CHANG R J, et al. Sapaugment: Learning a Sample Adaptive Policy for Data Augmentation[C]// Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2021:4040-4044.
[30] FAN C, DING M, YI J, et al. Two-stage Deep Spectrum Fusion for Noise-Robust End-To-End Speech Recognition[J]. Applied Acoustics, 2023, 212:1-10.

[31] SUN L, YOLWAS N, JIANG L. A Method Improves Speech Recognition with Contrastive Learning in Low-Resource Languages[J]. Applied Sciences, 2023, 13(8): 1-14.

[32] YANG H, ZHANG M, TAO S, et al. Chinese ASR and NER Improvement Based on Whisper Fine-Tuning[C]// Proceedings of 2023 25th International Conference on Advanced Communication Technology. IEEE, 2023: 213-217.

[33] GAO H, WANG X, KANG S, et al. Seamless Equal Accuracy Ratio for Inclusive CTC Speech Recognition[J]. Speech Communication: An International Journal, 2022, 136: 76-83.

[34] MIRKO A, SIMONE B, LUIGI C, et al. Semi-Supervised Cross-Lingual Speech Emotion Recognition[J]. Expert Systems With Applications, 2024, 237(A): 1-11.

[35] SHI X, LIU X, XU C, et al. Cross-Lingual Offensive Speech Identificationwith Transfer Learning for Low-Resource Languages[J]. Computers and Electrical Engineering, 2022, 101: 1-10.

[36] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]// Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM, 2009: 41-48.

[37] XU C, HU B, JIANG Y, et al. Dynamic Curriculum Learning for Low-resource Neural Machine Translation[C]// Proceedings of the 28th International Conference on Computational Linguistics. ACM, 2020: 3977-3989.

[38] KARAKASIDIS G, GRÓSZ T, KURIMO M. Comparison and Analysis of New Curriculum Criteria for End-to-End ASR[C]// Proceedings of INTERSPEECH 2022. 2022: 66-70.

[39] KUZNETSOVA A, KUMAR A, FOX J D, et al. Curriculum Optimization for Low-Resource Speech Recognition[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2022: 8187-8191.

[40] QIAN Y, ZHOU Z. Optimizing Data Usage for Low-Resource Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 394-403.

[41] ZHOU Z, WANG W, ZHANG W, et al. Exploring Effective Data Utilization for Low-Resource Speech Recognition[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2022: 8192-8196.

[42] MOHAMMADI S H, KAIN A. An Overviewof Voice Conversion Systems[J]. Speech Communication, 2017, 88: 65-82.

[43] KANEKO T, KAMEOKA H. CycleGANVC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks[C]// Proceedings of the 26th European Signal Processing Conference. Piscataway, NJ: IEEE, 2018: 2100-2104.

[44] ZHANG J X, LING Z H, LIU L J, et al. Sequence-to-sequence Acoustic Modeling for Voice Conversion[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 631-644.

[45] LI J Y, TU W P, XIAO L. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion[C]// Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2023: 1-5.

[46] THIENPONDT J, DEMUYNCK K. Transfer Learning for Robust Low-Resource Children's Speech ASR with Transformers and Source-Filter Warping[C]// Proceedings of INTERSPEECH 2022. 2022: 2213-2217.

[47] AZIZAH K, JATMIKO W. Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages[J]. IEEE Access, 2022, 10: 5895-5911.

[48] HALPERN B, FENG S, SON R V, et al. Low-resource Automatic Speech Recognitionand Error Analyses of Oral Cancer Speech[J]. Speech Communication, 2022, 141: 14-27.

[49] QIN S Q. Research on Modeling Unit for Tibetan Speech Recognition[D]. Tianjin: Tianjin University, 2022.

[50] QIN S Q, WANG L B, LI S, et al. Improving Low-resource Tibetan End-to-end ASR by Multilingual and Multilevel Unit Modeling[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2022, 2: 1-10.

[51] SOKY K, GONG Z, LI S. Nict-Tib1: A Public Speech Corpusof Lhasa Dialect for Benchmarking Tibetan Language Speech Recognition Systems[C]// Proceedings of 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques. IEEE, 2022: 1-5.

[52] ANOOP C S, RAMAKRISHNAN A G. Suitability of Syllable-Based Modeling Units for End-To-End Speech Recognition in Sanskrit and Other Indian Languages[J]. Expert Systems with Application, 2023, 220: 1-9.

[53] SHETTY V M, METILDA S M N J, UMESH S. Investigation of Speaker-adaptation methods in Transformer based ASR[J]. arXiv: 2008. 03247, 2020.

[54] NING J W. Research on End-To-End Semi-Supervised Speech Recognition Under Low-Resource Condition[D]. Lanzhou: Northwest Minzu University, 2023.

[55] QIAN Y, LIU J. Articulatory Feature Based Multilingual MLPS for Low-Resource Speech Recognition[C]// Proceedings of INTERSPEECH 2012. 2012: 2602-2605.

[56] LI S, DING C, LU X, et al. End-toend Articulatory Attribute Modeling for Low-Resource Multilingual Speech Recognition[C]// Proceedings of INTERSPEECH 2019. 2019: 2145-2149.

[57] SUBI A. Research on Uyghur Speech Recognition Based on End-to-End Modeling[D]. Urumqi: Xinjiang University, 2021.

[58] SHEN Z J, GUO W. Vietnamese Speech Recognition Based on Pre-training and Phone-Based Byte-Pair Encoding[J]. Journal of Data Acquisition and Processing, 2023, 38(1): 101-110.

[59] BALÁZS T, GYRGY S, TIBOR F, et al. Deep Transformer Based Data Augmentation with Subword Units for Morphologically Rich Online ASR[J]. arXiv: 2007. 06949, 2020.

[60] SABRINA J M, JASON E. Spell Once, Summon Anywhere: A Two-Level Open-Vocabulary Language Model[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. ACM, 2018: 6843-6850.

[61] EGOROVA E, VYDANA H K, BURGET L, et al. Spelling-Aware Word-Based End-to-End ASR[J]. IEEE Signal Processing Letters, 2022, 29: 1729-1733.

[62] DIWAN A, JYOTHI P. Reduce and Reconstruct: ASR for Low-Resource Phonetic Languages[C]// Proceedings of INTER-

SPEECH 2021. 2021:3445-3449.
[63] CHUNG H, LI J, LIU P F, et al. Improving Rare Words Recognition through Homophone Extension and Unified Writing for Low-resource Cantonese Speech Recognition[C]// Proceedings of the 13th International Symposium on Chinese Spoken Language Processing. IEEE, 2022:26-30.
[64] CHUNGY, ZHANG Y, HAN W, et al. w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training[C]// Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop. IEEE, 2021:244-250.
[65] ANOOPC S, RAMAKRISHNAN A G. Exploring a Unified ASR for Multiple South Indian Languages Leveraging Multilingual Acoustic and Language Models[C]// Proceedings of 2022 IEEE Spoken Language Technology Workshop. IEEE, 2023: 830-837.
[66] MORIYA T, SATO H, TANAKA T, et al. Self-Distilling Attention Weights For CTC-Based ASR Systems[C]// Proceedings of INTERSPEECH 2020. ISCA, 2020:6894-6898.
[67] MORIYA T, OCHIAI T, KARITA S, et al. Self-Distillation for Improving CTC-Transformer-Based ASR Systems[C]// Proceedings of INTERSPEECH 2020. ISCA, 2020:546-550.
[68] AHMAD R, JALAL M A, FAROOQ M U, et al. Towards Domain Generalisation in ASR with Elitist Sampling and Ensemble Knowledge Distillation[C]// Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2023:1-5.
[69] ŻELASKO P, FENG S, VELÁZQUEZ L M, et al. Discovering Phonetic Inventories with Crosslingual Automatic Speech Recognition[J]. Computer Speech & Language, 2022, 74:1-23.
[70] ZENG Z, PHAM V T, XU H, et al. Leveraging Text Data Using Hybrid Transformer-LSTM Based End-to-End ASR in Transfer Learning[C]// Proceedings of the 12th International Symposium on Chinese Spoken Language Processing. IEEE, 2021:1-5.
[71] YUSUF B, GANDHE A, SOKOLOV A. Usted: Improving ASR with a Unified Speech and Text Encoder-Decoder[C]// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2022:8297-8301.
[72] SUN E, LI J, XUE J, et al. Pre-training End-to-end ASR Models with Augmented Speech Samples Queried by Text[J]. arXiv: 2307.16332, 2023.
[73] LU X, SHEN P, YU T. Hierarchical Cross-Modality Knowledge Transfer with Sinkhorn Attention for CTC-Based ASR[C]// Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2024:13116-13120.
[74] LAPTEV A, KOROSTIK R, SVISCHEV A, et al. You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation[C]// Proceedings of the 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. IEEE, 2020:439-444.
[75] QU L, WEBER C, WERMTER S. Emphasizing Unseen Words: New Vocabulary Acquisition for End-To-End Speech Recognition[J]. Neural Networks, 2023, 161:494-504.
[76] HEIGOLD G, MORENO I, BENGIO S, et al. End-to-end Text-Dependent Speaker Verification[C]// Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2016:5115-5119.
[77] NATHANIEL R, PEREZ O, SWETHA G, et al. When Is TTS Augmentation Through a Pivot Language Useful? [C]// Proceedings of INTERSPEECH 2022. ISCA, 2022:3538-3542.
[78] MI C, XIE L, ZHANG Y. Improving Data Augmentation for Low Resource Speech-To-Text Translation with Diverse Paraphrasing[J]. Neural Networks, 2022, 148:194-205.
[79] SOKY K, LI S, MIMURA M, et al. Leveraging Simultaneous Translation for Enhancing Transcription of Low-resource Language via Cross Attention Mechanism[C]// Proceedings of INTERSPEECH 2022. ISCA, 2022:1362-1366.
[80] WANG J, ZHU Y, FAN R, et al. Low Resource German ASR with Untranscribed Data Spoken by Non-Native Children[C]// Proceedings of INTERSPEECH 2021. ISCA, 2021:1279-1283.
[81] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations [C]// Proceedings of the International Conference on Machine Learning. ACM, 2020:1597-1607.
[82] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020:9729-9738.
[83] LIN C E, CHEN K Y. A Lexical-aware Non-autoregressive Transformer-based ASR Model[C]// Proceedings of INTERSPEECH 2023. ISCA, 2023:1434-1438.
[84] BERREBBI D, SHI J, YAN B, et al. Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation[C]// Proceedings of INTERSPEECH 2022. ISCA, 2022:3533-3537.
[85] ZATVORNITSKIY A. Low-Cost Training of Speech Recognition System for Hindi ASR Challenge 2022[C]// Proceedings of SPECOM 2022. Cham: Springer, 2022, 13721:712-718.
[86] HAIDAR M A, XING C, REZAGHOLIZADEH M. Transformer-Based ASR Incorporating Time-Reduction Layer and Fine-Tuning with Self-Knowledge Distillation[C]// Proceedings of INTERSPEECH 2021. ISCA, 2021:2102-2106.
[87] ZHU W, JIN H, CHEN J, et al. A Hybrid Acoustic Model Based on PDP Coding For Resolving Articulation Differences in Low-Resource Speech Recognition[J]. Applied Acoustics, 2022, 192: 1-11.
[88] AMBUJ M, NAVONIL M, RISHABH B, et al. A Review of Deep Learning Techniques for Speech Processing[J]. Information Fusion, 2023, 99:1-55.
[89] YADAV H, GUPTA A, RALLABANDI S K, et al. Intent Classification Using Pre-Trained Language Agnostic Embeddings For Low Resource Languages[C]// Proceedings of INTERSPEECH 2022. ISCA, 2022:3473-3477.
[90] XIANG Z H, GU X, RAO C Z, et al. Research on Low-resource Qingdao Dialect Speech Recognition Method[J]. Computer Technology and Development, 2024, 34(4):146-152.
[91] LEEC Y, GLASS J. A Nonparametric Bayesian Approach to Acoustic Model Discovery[C]// Proceedings of the 50th Annual

Meeting of the Association for Computational Linguistics. ACL, 2012:40-49.

[92] WU B,SAKTI S,NAKAMURA S. Incorporating Discriminative DPGMM Posteriorgrams for Low-Resource ASR[C]// Proceedings of 2021 IEEE Spoken Language Technology Workshop. IEEE,2021:201-208.

[93] WU B,SAKTI S,ZHANG J,et al. Modeling Unsupervised Empirical Adaptation by DPGMM and DPGMM-RNN Hybrid Model to Extract Perceptual Features for Low-Resource ASR [J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2022,30:901-916.

[94] XU L,ZHAO Y,XU X,et al. Latent Regression Bayesian Network for Speech Representation[J]. Electronics,2023,12(15): 1-12.

[95] KRISHNA V,SAI T,GANAPATHY S. Representation Learning With Hidden Unit Clustering for Low Resource Speech Applications[J]. IEEE/ACM Transactions on Audio,Speech, and Language Processing,2024,32:1036-1047.

[96] LI X,DALMIA S,LI J,et al. Universal Phone Recognition with a Multilingual Allophone System[C]// Proceedings of 2020 IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2020:8249-8253.

[97] HARDIK S,KIRAN P T,VIKAS A,et al. SRI-B End-to-End System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages[C]// Proceedings of INTERSPEECH 2021. ISCA,2021:2456-2460.

[98] HUANG R. Integrating Categorical Features in End-To-End ASR[J]. arXiv:2110.03047,2021.

[99] CHEN Z,LI J,LIU H,et al. Learning Multi-Scale Features for Speech Emotion Recognition with Connection Attention Mechanism[J]. Expert Systems with Applications,2023,214:1-10.

[100] LIU M,ALEX N J R,VIJAYARAJAN R,et al. Multiscale-multichannel Feature Extraction and Classification through One-Dimensional Convolutional Neural Network for Speech Emotion Recognition[J]. Speech Communication,2024,156:1-14.

[101] SUNDARARAMAN M N,KUMAR A,VEPA J. PhonemeBERT: Joint Language Modelling of Phoneme Sequence and ASR Transcript[C]// Proceedings of INTERSPEECH 2021. ISCA,2021:3236-3240.

[102] TSAI H S,CHANG H J,HUANG W C,et al. SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. ACL,2022:8479-8492.

[103] DOSSOU B,TONJA A L,EMEZUE C,et al. Adapting Pretrained ASR Models to Low-resource Clinical Speech using Epistemic Uncertainty-based Data Selection[J]. arXiv: 2306.02105,2023.

[104] FAROOQ M U,NARAYANA D A H,HAIN T. Non-Linear Pairwise Language Mappings for Low-Resource Multilingual Acoustic Model Fusion[C]// Proceedings of INTERSPEECH 2022. ISCA,2022:4850-4854.

[105] LIN Y Q,DANG J W,WANG L B,et al. Disordered Speech Recognition Considering Low Resources and Abnormal Articulation[J]. Speech Communication,2023,155:1-9.

[106] LYU Y,WANG L,GE M,et al. Compressing Transformer-Based ASR Model by Task-Driven Loss and Attention-Based Multi-Level Feature Distillation[C]// Proceedings of 2022 IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2022:7992-7996.

[107] KAK S,ZHUO G,SHENG L. Nict-Tib1:A Public Speech Corpusof Lhasa Dialect for Benchmarking Tibetan Language Speech Recognition Systems[C]// Proceedings of the 25th Conference of the Oriental COCOSDA. IEEE,2022:1-5.

[108] YANG J,LI H G,ZHANG X L. The Research on the Construction of Bai Language Speech Corpus[J]. Journal of Dali University,2017,2(12):21-26.

[109] REN P,XIAO Y,CHANG X,et al. A Survey of Deep Active Learning[J]. ACM Computing Surveys,2021,54(9):1-40.

[110] SEVERIN I,PUSKURUK,CEDRIC J,et al. Universal Data Tool[EB/OL]. (2021-02-25)[2024-01-15]. https://github.com/UniversalDataTool/universal-data-tool.

[111] NOVAK J R,MINEMATSU N,HIROSE K. WFST-based Grapheme-To-Phoneme Conversion:Open Source Tools For Alignment,Model-Building And Decoding[C]// Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing. ACM,2012:45-49.

[112] MORTENSEN D R,DALMIA S,LITTELL P. Epitran:Precision G2P for many languages[C]// Proceedings of the 7th International Conference on Language Resources and Evaluation. ACM,2018:2710-2714.

[113] HAN X,WANG Y T,FENG J L,et al. A Survey of Transformer-Based Multimodal Pre-Trained Modals[J]. Neurocomputing, 2023,515:89-106.

YANG Jian, born in 1976, Ph.D, associate professor, is a member of CCF (No. 14480M). His main research interests include speech recognition and deep learning.

(责任编辑:何杨)