中国方言语音识别技术的发展概述

袁稳沉[1*] 方明康[2] 单越[2] 杨安[1]

1 杭州职业技术学院，浙江省杭州市，310020；

2 美欣达集团有限公司，浙江省湖州市，313000；

摘要：本文讨论了适老化语音识别面临的独特挑战，如老年人多使用方言，以及老年人语音特征的变化对数据需求的影响;综述了近年来自动语音识别技术的主要发展，包括传统的高斯混合模型-隐马尔可夫模型、基于深度神经网络的混合模型，以及端到端方法如 CTC-注意力机制和 Transformer 架构的进展,及其在中国方言语音识别中的应用。

关键词：适老化语音识别；卷积神经网络；端到端模型

DOI：10.69979/3041-0673.24.11.021

随着全球人口老龄化的加剧，老年人群体对于智能科技的需求日益增长，开发和优化适老化方言语音识别算法对于提高老年人生活质量、促进数字包容性具有重要意义。在中国，方言的多样性为老年人使用智能语音识别技术带来了挑战。当前的语音识别系统大多基于普通话或标准语言训练，对于方言的识别能力有限，这限制了老年人使用智能设备的便利性。本文拟梳理国内外关于自动语言识别（language identification， LID）技术的发展历程与现状。

自动语言识别技术的目标是在无需人工干预的情况下确定给定语音片段的语言信息。信号处理、模式识别、认知科学和机器学习等领域的技术突破推动了 LID 系统的发展[1]。早期语种识别技术常使用高斯混合模型（Gaussian Mixture Models，GMM），该模型假设每一种语言的声学特征由一个 GMM 生成，通过比较不同 GMM 输出的概率来识别语种。GMM 因其能够近似复杂数据分布和模拟语言的底层声学类别而成为LID中的流行选择[2]。然而 GMM 模型对说话人和信道变化等非语言效应敏感，需要大量训练数据[3]。为克服这一困难，研究者提出了最大互信息训练（Maximum Mutual Information，MMI），MMI 训练是一种判别性方法，它最大化了正确识别的后验概率，尤其在特征空间中类别高度重叠的情况下优势明显[4]。为了判别相似语音信号，研究者提出了异方差线性判别分析（Heteroscedastic Linear Discriminant Analysis，HLDA）方法，主要用于说话人识别。HLDA 用于去相关和降维特征，同时保留特征的判别能力。它在 LID 系统中一致性地提高了识别性能[5]。另外，隐马尔可夫模型（Hidden Markov Models，HMM）通过在多个状态间模拟语音特征的动态，为 GMM 提供了扩展，从而在 LID 性能上提供了额外的改进[6]。

从上世纪 80 年代起，高斯混合模型-隐马尔可夫模型(GMM-HMM)一直主导了语音识别的研究。然而 GMM 本身的表现更多依赖于声学特征的质量。近年来，深度神经网络（Deep Neural Network，DNN）在语音识别领域的应用取得了显著进展，发展出了深度神经网络-隐马尔可夫模型(DNN-HMM)。此类混合模型利用 HMM 处理序列数据的优势，同时发挥DNN在特征学习中的强大能力。在多语言数据集（如 GlobalPhone）上的实验表明，DNN-HMM 框架在跨语言任务中实现了低词错误率，在语音识别精度上表现优异[7, 8]。但是较少标注数据会导致其性能下降。为次，研究人员提出了迁移学习、预训练模型等方法[9, 10]。

端到端(End-to-End)语音识别模型是另一种主流语音识别架构，其与混合模型有不同的理念。这些模型通常采用编码器-解码器架构，通过直接学习从语音到文本的映射，简化了传统的语音识别流水线。Bahdanau 等人（2016）提出了基于注意力机制（Attention Mechanism）的端到端大词汇量语音识别，这种机制可以有选择地关注信息的不同方面[11]。端到端语音识别模型结合注意力机制，能显著提高语音识别的效率。例如，在 LibriSpeech 数据集上，基于长短时记忆网络（Long Short-Term Memory，LSTM）和注意力机制的模型在“test-clean”数据集上的单词错误率低至 2.44%，表明其在干净环境下的性能接近最先进水平[8]。Kim 等人（2017）结合了 CTC 和注意力机制的优点，提出了一种新的端到端框架，即混合 CTC/注意力机制，通过多目标学习框架提高模型的鲁棒性。端到端方法直接将语音序列映射到文本序列，简化了复杂的建模过程[12]。然而，端到端方法的鲁棒性在低资源或嘈杂环境中可能不如混合模型[13]。在标注数据有限的情况下，数据增强、迁移学习以及生成对抗网络（Generative Adversarial Network，GAN）等方法被广泛应用。这些技术通过生成合成

数据或迁移相关领域的知识，增强了模型的鲁棒性。例如，使用条件 GAN 生成新的语音样本有效平衡了数据集，从而提高了模型性能[14]。

深度学习的进一步发展， 一些高级构架正在打破混合模型和端到端方法的界限。如 Transformer 的架构，它既可以融入混合模型的设计，也可以作为端到端方法的核心组件。这使模型能够捕捉语音数据中的长程依赖关系，使其在嘈杂或复杂环境中的表现优于传统方法[7]。例如 Transformer-HMM 模型结合 Transformer 的强大特征提取能力与 HMM 的时序建模特性。另外，有大量基于 Transformer 的端到端模型，如 Speech-Transformer[15, 16]、多任务端到端 Transformer（通过联合优化语音识别和其他语音任务，进一步提升泛化能力）[14]。Transformer 作为现代深度学习的基础模块，为混合模型提供增强能力，同时推动端到端方法向更高效和更鲁棒的方向发展。另外，一种基于 Transformer 构架的变体 Conformer（卷积增强 Transformer）结合了卷积神经网络（CNN）和 Transformer 的优点，专为序列数据的处理而设计，特别适用于自动语音识别（ASR）任务[17]。Conformer 结合了 Transformer 的自注意力机制和卷积层的优势：自注意力机制擅长捕获输入序列中的全局依赖；卷积层则能够提取局部上下文信息，尤其适合捕捉语音信号中的细微时序特征。这种结合解决了传统 Transformer 在捕获语音局部特征时的不足。

中国方言语音识别技术经历了类似的发展。Juang 和 Rabiner（ 1991）将 HMM 作为中文方言语音识别的主要方法，因其准确的数学模型能够从语音数据中计算出训练模型参数[18]。Reynolds（2009）展示了 GMM 作为参数模型，结合 HMM 使用，能够构建中文方言语音识别系统[19]。早期 HMM 系统在方言适配上的成功更多依赖于特定领域的知识和数据特性。随着深度学习技术的发展，基于深度神经网络的方法已成为处理复杂语音模式的强大替代方案。Pan 等人（2012）研究了 DNN 在大词汇量连续语音识别中的应用，表明 DNN 在声学建模方面超越了 GMM[20]。O’Shea 和 Nash（2015）介绍了 CNN 在处理时间序列数据方面的优势，特别是在语音识别中的表现[21]。中国方言语音识别面临着诸多挑战，如数据收集难度较大导致训练样本稀缺，以及方言间在音系学和语法上的显著差异。为应对这些问题，一些研究结合了传统 HMM 在时序建模上的优势与 DNN 的特征学习能力，即采用 DNN-HMM 混合模型[22, 23]。另一技术路线-端到端语音识别模型也在中文语音识别领域得到了广泛的应用[21]。Hori 等人（2017）提出了连接时序分类（Connectionist Temporal Classification，CTC）能够解决时间数据分类任务，通过神经网络计算输入数据和给定输出之间的误差[24]。Chen 等人（2024）提出了一个名为 Qifusion-Net 的层适应融合（Layer-adapted Fusion，LAF）模型，不同于传统端到端方法，该模型不需要关于目标口音的任何先验知识，基于动态块策略，能够实现流式解码并提取帧级声学特征，有助于细粒度信息融合。实验结果表明，该方法在 KeSpeech 和包含 6 个不同地域（四川、山西、山东、江苏、湖南、广东）的 MagicData-RMAC 多口音测试数据集上相对于基线模型在字符错误率上分别降低了 22.1%和 17.2%[25]。

中国方言的语音识别技术得到了长足的发展，然而目前尚未见太多适老化的方言语音识别模型研究。适老化语音识别模型须捕捉老年人的语音特征变化，须针对性地调整算法以适应老年人的语音特性，包括调整网络结构和增强模型的泛化能力，以期开发用户自适应机制，使系统能够根据个体差异（地区、年龄）进行动态调整，为后续的智能语音应用、适老化智能设备人机交互模块的开发建立基础。

参考文献

[1]Li, H., B. Ma, and K.A. Lee, Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE, 2013. 101(5): p. 1136-1159.

[2]Zissman, M.A. Automatic language identification using Gaussian mixture and hidden Markov models. in 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing. 1993. IEEE.

[3]Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, Speaker verification using adapted Gaussian mixture models. Digital signal processing, 2000. 10(1-3): p. 19-41.

[3]Burget, L., P. Matejka, and J. Cernocky. Discriminative training techniques for acoustic language identification. in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. 2006. IEEE.

[4]Tipping, M.E. and C.M. Bishop, Mixtures of probabilistic principal component analyzers. Neural computation, 1999. 11(2): p. 443-482.

作者简介：袁稳沉，1991.06，男，汉，浙江，博士，教师，讲师，人工智能、计算力学，杭州职业技术学院。