语音识别技术及应用综述

■ 涂 冲 金利英 王中任 刘海生 邹雨杰

语音识别技术是将人类的语音信号转换为可读文本或命令的过程。语音识别技术发展历程长，从早期特定人、孤立词识别发展到当下多人、连续及多语言混合识别，精度与效率显著提升。本文将对语音识别技术的发展历程、关键技术、应用现状、未来挑战以及未来趋势进行综述。

一、引言

在人类的交流体系中，语音是最为自然且高效的方式之一。随着科技的不断进步，人机交互成为拓展人类与技术互动边界的重要领域，语音识别技术应运而生，成为该领域的核心支撑。语音识别技术的发展历程漫长且充满突破，如今，它已从最初的概念验证阶段，迈入广泛应用的成熟时期。尽管语音识别技术已取得斐然成绩，但距离完美的人机交互仍有差距，如噪声干扰、口音方言、有限语义理解等。在这一背景下，全面梳理语音识别技术的发展脉络、关键技术、应用现状与挑战，展望未来发展趋势，对推动该技术持续创新具有重要的理论与实践意义。

二、发展历程

语音识别技术的发展可以追溯到 20 世纪 50 年代的早期探索阶段，贝尔实验室在 1952 年研制出能识别 10 个英文数字的 Audry 系统，该系统把共振峰定位到每个单词的功率谱中进行辨识，开启语音识别研究序幕 [1]。

20 世纪 70 年代，语音识别进入快速发展阶段。随着最大似然线性回归和最大后验概率估计的使用，解决了隐马尔可夫模型（HMM）自适应问题，后使用高斯混分模型（GMM）用于统计观测概率取得不错成果，GMM-HMM 和 N-Gram 在 20 世纪末成为主流 [2]。

进入 21 世纪，语音识别进入深度发展阶段。随着计算机能力以及神经网络算法有较大发展，基于深度学习的方法得到广泛应用 [3]。如深度神经网络（DNN）、卷积神经网络（CNN）、循环神经网络（RNN）及其变体长短时记忆网络（LSTM）和门控循环单元（GRU）等。

三、关键技术

（一）预处理

在对语音进行识别之前，对采集的语音进行处理。目的是将声音从模拟信号转为数字信号，便于计算机的存储与识别，抗干扰能力强。模拟信号，可以反映声波频率和振幅的变化，之后由 ADC 通过采样、量化、编码这 3 个步骤转化为数字信号以便计算机处理 [4]。

预处理是语音识别的基础，决定着语音识别过程能否顺利进行，主要有预加重、分帧、加窗、端点检测等步骤 [5]。

端点检测：其目的是通过某种方式检测出有效语音段。如基于能量的方法、基于双门限方法。

预加重：本质是对高频及产生进行提高和补偿以提高后续识别准确性。

分帧加窗：语音信号是非平稳信号，但在极短时间内可视为平稳信号，所以可进行分帧操作。

（二）特征提取

语音信号的特征提取目的是将原始语音信号转换为能代表语音本质的参数过程，达到降维和除去冗余信息的效果。

常用的语音特征有滤波器组（FBANK 特征）、梅尔频率倒谱系数（MFCC）、相对谱变换与感知线性预测等，为了获得更多的特征作为声学模型的输入可用线性判别分析（LDA）和最大似然线性变换（MLLT）方法 [6]。

因 MFCC 可很好地模拟人耳特性，所以应用广泛。MFCC 将声谱图映射到梅尔频率尺度上再进行离散余弦变换（DCT）得到频带系数即

MFCC 系数，LPCC 与 MFCC 相似但处理高频分量时更准确 [7]。

（三）声学模型

声学模型是语音识别系统的核心，负责将特征向量映射到音素或字的序列。传统方法中，HMM 被广泛使用。近年来，DNN、CNN、端到端语音识别模型和 RNN 等深度学习模型在声学建模方面取得了突破性进展。

1. DNN

一种基于神经网络的机器学习技术，在语音识别等多个领域常见。可以自动地从大量语音数据中学习相应语音特征，包含时频及上下文信息等。可与其他语音识别技术相结合，常见的有 DNN-HMM。文献 [8] 中使用相对传统声学模型（如 HMM、GMM 等），可利用帧上下文信息和建模能力更强的 DNN 来进行语音识别，针对 DNN 无法对语音长时相关性建模而提出 TDNN，用 MLE 对 HMM 和 DNN 进行训练的同时用 DT 对 DNN 和 TDNN 进行训练，经测试可得 TDNN 优于 DNN 优于 HMM。

2. CNN

一种深度学习模型，适用于处理具有网格结构的数据，在语音识别和图像识别领域较常见，由卷积层、激活函数层、池化层、全连接层构成。可以从大量数据中自动提取语音特征。但所消耗计算资源较高，对于处理长序列依赖关系时能力有限。文献 [9] 中提出 CNN 对局部特征进行提取的同时运用多头注意力机制捕获全局信息来解决长序列依赖问题，其中针对多头注意力机制于子空间维度冲突等问题，可在此基础上利用膨胀卷积和多分支结构，即因每条分支的信息量减少使得模型可使用较多注意力头数。

3. HMM

一种常用的统计学模型，具有马尔可夫性。HMM 实时处理能力有限，并且难处理复杂模型，在实际应用中，通常将 HMM 与其他技术相结合，如 GMM-HMM、DNN-HMM。文献 [10] 中，分别对应用了 MFCC 的 GMM-HMM 和 Fbank 的 DNN-HMM 进行模型训练和测试方言识别，最后 DNN-HMM 的识别准确性较优于 GMM-HMM。

4. 端到端语音识别模型

与传统语音识别方法不同，端到端语音识别模型无须进行特征提取及声学建模，而是直接将语音信号转为文本。该模型架构较简单，可更好地学习上下文信息。在文献 [11] 中，CNN 注重局部关联信息，而 RNN 对于处理长序列依赖时具有很好的效果，基于二者优点所构建的 CRNN-CTC 模型，对模型进行相关指标测试如 FRR、FAR 取得较好的效果。对于目前鲁棒性改进缺乏知识数据有所帮助。

5. RNN

在处理序列类型的语音数据时有着不错的效果，如理解上下文语义。但在实际应用中，语音序列长度增加时容易产生梯度消失或者梯度爆炸，因此常常与其他模型结合来弥补缺点。针对梯度爆炸和梯度消失等问题，对 RNN 改进为 LSTM，其中通过输入、遗忘、输出门可防止梯度问题，可通过神经元之间的线性连接来构建深层次的 LSTM[12]。

（四）语言模型

语言模型用于评估一个词序列出现的概率，它对提高语音识别的准确性至关重要。主要有统计语言模型 N-Gram、基于神经网络的语言模型，预训练语言模型也常被运用。

1. 预训练语言模型

在大量文本数据上进行预训练从而学习语言的通用特征表示，之后使用标注数据进行微调，较常见的模型如 BERT，常用于对语音识别后的文本进行校正。文献 [13] 对 BERT 进行改进，先检错然后 MACBERT 纠错，即针对错字与原字相近的特性设计出“置信度 – 相似度”指标并对相似度计算进行改进，对于多音错误文本数据较少的情况提出基于 BERT 优化模型，通过两种方法对文本进行增强的同时也对文本进行纠错，在数据集 Thchs-30 和 NLPCC2018 上取得不错的效果。

2. N-Gram 和神经网络

N-Gram 语言模型认为某词的出现与前面词有关，通过 N-Gram 出现频率来计算文本序列概率，虽然计算简单易实现，但处理长期依赖关系的能力有限，因此常与神经网络相结合以更好地理解文本语义。在实际应用中将二者结合，首先通过 N-Gram 将候选词范围缩小，再通过神经网络进行精准预测，之后可运用迁移学习进行优化，使其能对其他语言进行预测 [14]。

（五）解码算法

解码算法用于在声学模型和语言模型的基础上寻找最可能的词序列。波束搜索（Beam Search）是目前最常用的解码算法之一，它通过限制搜索空间来平衡识别的准确性和计算效率。除此之外还有

CTC 解码、维比特算法、WFST 算法。

1. WFST

将声学模型、语言模型、词典整合到有限状态自动机中。状态间可通过边来转移。适用于大规模语音识别系统，且支持各种语言模型，如 N-Gram 和神经网络语言模型。

2. CTC 解码

该方式对输出序列中插入空白标签来实现特征序列与文本序列的对齐，因无须人工标注而在端到端语音识别系统中应用广泛。相对于波束搜索算法，计算速度快，但因波束搜索算法保留多个候选路径使得在复杂语音识别时可以找到全局最优路径。在文献 [15] 中，使用了包含 CTC 贪婪搜索、CTC 前缀波束搜索在内的 4 种算法，其中 CTC 贪婪搜索解码时选概率最大的输出，有效地提高了最优解。

四、结语

语音识别技术作为人工智能领域的重要分支，已经取得了显著的进展，并在多个领域得到了广泛应用。随着算法和硬件技术的进一步发展，语音识别技术将更加智能化、个性化。然而，语音识别技术仍面临挑战。未来，随着深度学习、迁移学习等技术的持续创新，有望在模型优化、特征提取等方面取得突破，提升识别准确率和泛化能力。同时，与自然语言处理、计算机视觉等多技术融合，将推动语音识别技术迈向人机深度交互的新阶段，拓展其在医疗、教育、金融等更多领域的应用，为社会发展带来更多可能。

引用

[1] 鹿哲源,牛小明,康林,等.人机交互语音识别发展及军事应用分析[J].兵工自动化,2023,42(4):21-25.

[2] 佚名.语音识别技术科普与发展历史[J].科技视界,2023(2):38-39.

[3] 程美,王力华.医疗智能语音技术与应用综述[J].中国数字医学,2021,16(8):1-7.

[4] 李雪莹.基于深度学习的语音识别技术研究[D].北京:北方工业大学,2024.

[5] 李龙.基于语言功能的智能交通机器人知识自主生成机制[D].兰州:兰州理工大学,2023.

[6] 黄志东.鲁棒性语音识别技术研究综述[J].信息通信,2019(11):20-22.

[7] 李焕贞,孙茜.基于语音识别技术的智能机器人控制系统设计与应用[J].无线互联科技,2023,20(15):41-44.

[8] 周婕.基于Kaldi的中文语音识别研究[D].南京:南京邮电大学,2022.

[9] 刘凯.基于语音信号的多模态情感识别关键技术研究[D].济南:山东大学,2023.

[10] 熊金准.用于襄阳方言语音识别的人机交互系统研究[D].襄阳:湖北文理学院,2023.

[11] 郑若伟.低资源端到端语音关键词检测技术研究及实现[D].广州:华南理工大学,2023.

[12] 叶硕,褚钰,王祎,等.语音识别中声学模型研究综述[J].计算机技术与发展,2020,30(3):181-186.

[13] 邢月晗.语音识别后的中文文本纠错系统设计与实现[D].北京:北京邮电大学,2023.

[14] 张祥.多语言语音识别技术在智能语音助手中的应用研究[J].电声技术,2024,48(4):42-44.

[15] 王超.基于深度学习的端到端藏语语音识别研究[D].拉萨:西藏大学,2023.

基金项目：弹射座椅姿态与轨迹控制技术研究（2021BID001）；基于视觉识别的四旋翼无人机自主降落研究（2024pygpzk05）；基于人工智能的智能驾驶系统优化研究（ZDSYS202406）

作者简介：涂冲，湖北文理学院，硕士研究生，研究方向为语音识别；王中任，湖北文理学院，博士，教授，研究方向为机器视觉、智能机器人；刘海生，湖北文理学院，本科，教授，研究方向为机械；邹雨杰，湖北文理学院，本科在读，研究方向为智能控制。

通讯作者：金利英，湖北文理学院，博士，副教授，研究方向为智能控制及算法、信号处理及故障诊断等。