智能语音技术在构音障碍方向的研究进展与趋势

□文 / 赵欣然[1,2#]，刘柏[1,3#]，刘小康[4]，吴锡欣[5]，燕楠[4]，王甦菁[1*]

（1. 中国科学院心理研究所，北京 100101；2. 江苏科技大学 计算机学院，江苏 镇江 212003；3. 长春大学 计算机科学技术学院，长春 130022；4. 中国科学院深圳先进技术研究院 先进技术集成所，广东 深圳 518055；5. 香港中文大学 系统工程与工程管理学系，香港 HKG）

摘要：构音障碍是一种由中枢或者外周神经系统受损引发的言语障碍，往往伴随着发音混乱、发音错误、声音忽大忽小，以及音调失常等情况，导致构音障碍者的语音很难被他人听清，从而极大地影响到他们与社会之间的交流。近年来，智能语音技术慢慢走进了人们的生活。怎样利用智能语音技术助力构音障碍患者更好地融入社会，而不是造成更多的技术障碍，这应当是亟需解决的问题。本文首先论述构音障碍的疾病种类及其声学特征，其次阐述如何评定构音障碍的严重程度，再次介绍有关构音障碍语音识别、语音修复与语言合成的关键技术发展情况，针对标注数据缺乏、个体差异较大这些主要难点，提出自监督学习、多模态信息融合等解决办法，最后对未来智能语音技术在构音障碍领域的发展走向作出预测，期望给构音障碍领域的语音技术革新以及项目落地等提供一些理论支撑和参照。

关键词：智能语音技术；构音障碍；构音障碍语音识别；自监督预训练；多模态融合

中图分类号：TN912.34　**文献标志码：**A　**文章编号：**2096-5036(2025)05-0001-19

DOI：10.16453/j.2096-5036.202542

0 引言

沟通属于人类社会交往的基本需求，语言表达及理解能力会影响个体融入社会的水平。构音障碍（Dysarthria）是由神经肌肉功能出现异常引发构音器官活动受限所造成的言语障碍[1]。这类人群平时交流的时候常常会面临传达不畅，社交范围小等问题，极大地影响到自身的生命质量以及心理状态。

近年来，随着人工智能技术的快速发展，智能语音技术在多个领域取得突破性进展，在辅助构音障碍人群的交流方面也展现出巨大潜力。构音障碍相关语音技术的发展，使得含糊不清的语音能更好地被识别，从而帮助患者更高效地表达需求，改善生活自理能力并增加社交活动，为其生活质量带来根本性改善。本文

基金项目：国家自然科学基金（62276252，U23B2018，62271477）；深圳市重点基金项目（JCYJ20220818101411025，JCYJ20220818101217037）

将系统梳理智能语音技术在构音障碍领域的研究进展，探讨实际应用中的挑战及解决思路，为后续研究提供参考。

在康复训练方面，构音障碍者的语言康复往往耗时较长，且训练效果存在波动。即便坚持每日进行一小时的针对性训练，通常也需数周才能见效，并且极易受到个体状态变化的影响[2,3]。若智能语音识别技术能够实现对构音障碍语音的高准确率识别，患者便能够在康复期间借助该技术实现基本交流，从而减轻康复过程中的时间与经济压力，提升学习与工作的效率。

然而，智能语音技术在实际应用中仍面临诸多挑战：① 缺乏高质量的标注数据；② 患者语音与通用模型适配性不足，导致性能下降；③ 构音障碍语音在不同个体之间，甚至同一患者的不同情绪状态下，往往在发音方式、语速、声调等方面存在显著差异，进一步增加了模型的泛化难度。

尽管如此，相关研究已在语音检测、评估、合成与修复等方向取得显著进展。既往研究表明，构音障碍通常与神经系统损伤或发音器官受限密切相关，不同类型患者的语音表现差异显著。其声学特征常表现为语速减慢、音素替代、元音空间缩小和基频不稳定等[4–6]，这些发现为后续语音处理提供了理论依据。早期临床多依赖FDA-2、AIDS等主观量表评估病情，但结果受限于评定者差异，缺乏一致性[7]。因此，近年来研究逐渐转向结合声学特征与机器学习、深度学习方法进行客观评估，并在UASpeech、Nemours等数据库上取得了明显提升[8]。

在智能语音技术研究中，语音识别始终是构音障碍领域的核心任务。早期方法依赖MFCC、CQCC等手工特征及小规模语料，系统性能及跨人群的泛化能力有限。随着UASpeech、TORGO、CDSD等数据库的建立，基于深度学习的端到端识别逐渐成为主流，并结合自监督学习与多模态融合方法，有效缓解了数据不足与模型适配性差的问题[9]。与此同时，语音修复与合成也在快速发展，从统计参数合成到基于深度神经网络与扩散模型的方法，显著提升了语音的自然度与可懂度，并在一定程度上缓解了语料不足，为识别与评估提供了支持[10]。这些进展表明，智能语音技术正逐步从理论探索走向临床与应用，为改善构音障碍者的交流与康复提供可行路径。

与国际相比，国内相关研究起步较晚，但近年来发展迅速，并逐渐在中文韵律建模与本土化临床应用方面展现优势。在声学、语音和信号处理国际会议（ICASSP）以及国际言语交流协会年会（INTERSPEECH）这两个顶级会议上，有关构音障碍的智能语音技术论文数量从2016年起便呈稳定增长态势，早期近乎空白，之后每年都有几篇，这表明信号处理算法、深度学习，以及多模态融合等研究领域对于改善构音障碍语音识别的效果十分明显，相关论文数量情况可参照图1、图2。

智能语音技术对于助力构音障碍患者提高交流能力方面颇具意义，不过也遭遇诸多挑战，这些挑战涉及数据、建模、识别精度，以及实际适配性等方面，为应对这些问题，后面的章节将会从构音障碍的疾病种类、声学特性、严重程度评估、构音障碍语音识别、构音障碍语音修复和语音合成这几个方面展开论述，以探寻更具实用性、广泛性的技术途径。

1 构音障碍疾病种类

1.1 构音障碍的定义与生理基础

构音障碍主要表现为发声障碍（声带振动

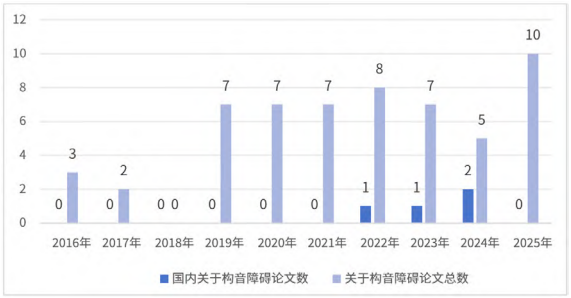

图 1 ICASSP 会议论文集上构音障碍论文数量

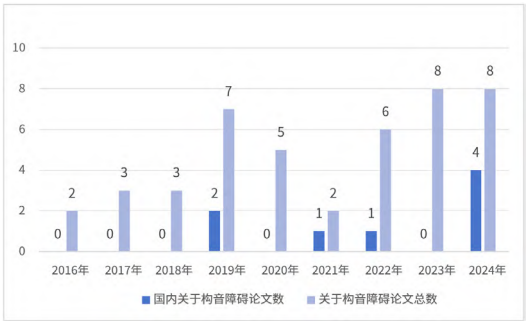

图 2 INTERSPEECH 会议论文集上构音障碍论文数量

异常导致的音质改变，如气声、嘶哑，以及音调、音量控制困难）、共鸣异常（软腭功能障碍引起的鼻音化或去鼻音化）、构音不清（舌、唇、下颌等构音器官运动障碍造成的辅音、元音发音错误或扭曲），以及韵律异常（言语节奏、重音模式、语调变化等超音段特征的改变，影响言语的自然流畅性）[11]。值得注意的是，构音障碍是发音障碍，但词义及语法通常保持正常。

1.2 构音障碍的常见病因及分类

神经解剖出现状况，或者外周构音结构受损，均有可能引发构音障碍，常见诱因涵盖脑血管疾病、颅脑外伤，以及部分神经系统疾病，如脑瘫及肌萎缩性侧索硬化，小脑病变和帕金森病等[12]，构音障碍可单独产生，亦可与失语症等其他语言障碍一同存在。

根据病因与临床表现，构音障碍的种类可以被分为运动性、器质性和功能性三大类[1]，如图 3 所示。

神经系统受损之后，会进一步影响到构音器官之间的协调运动能力，这是运动性构音障碍产生的关键因素，大致可以把这种类型区分为六种：痉挛型、弛缓型、共济失调型、运动过强型、运动过弱型，以及混合型。痉挛型、弛缓型均源于运动神经受损，其中痉挛型由于

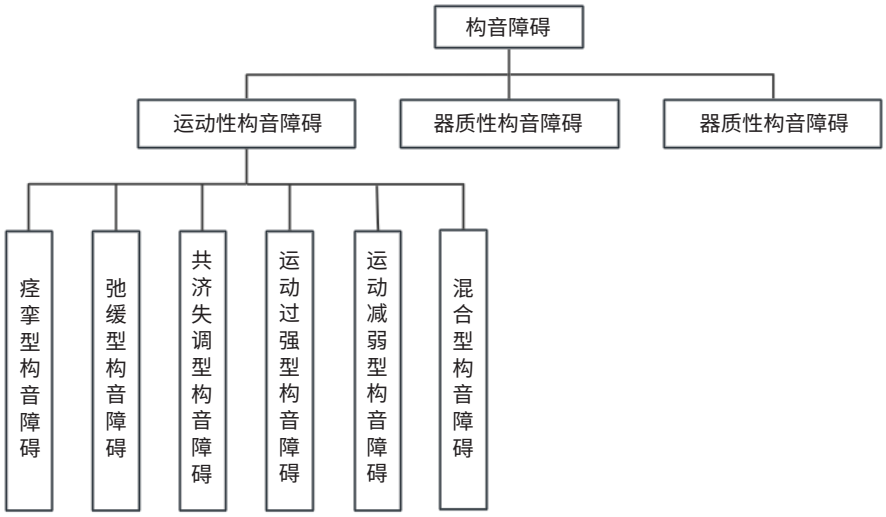

图 3 构音障碍疾病种类

表 1 健康对照组与构音障碍者的声学差异分析

声学特征	健康对照组表现	构音障碍者表现
语速与时长	语速正常，元音及音节时长符合语言规范，停顿规律有序	语速明显偏慢，元音持续时间延长，语音节奏不规则，存在异常停顿或拖延
声学与频谱稳定性	声学方差小，频谱特征（如梅尔频率倒谱系数）稳定，能量分布均衡，频谱倾斜度正常	声学方差大，频谱特征异常且能量分布不均，高频能量减少，频谱倾斜异常
发音准确性	爆破音等辅音误读率低，极少出现音素替换或丢失，擦音和塞擦音发音完整	爆破音误读率高，多为清浊音替换，频繁出现音素替换、丢失，常缺失词末擦音和词首塞擦音
元音空间特征	元音空间区域较大，元音离散度高，共振峰过渡正常	元音空间区域缩小，元音集中化，元音离散度低，共振峰过渡异常
基频特征	基频稳定，无抖动，音高变化丰富自然，极少出现音高中断	基频不稳定，部分存在抖动，音高多呈单调性或不规则波动，音高中断现象常见
发声稳定性	声带振动稳定，Jitter、Shimmer 值低，谐波噪声比高，噪声少	声带振动不稳定，Jitter、Shimmer 值高，谐噪比低，噪声成分多，语音质量下降
共振峰特征	共振峰范围及 F2/F1 范围稳定，不随语境等因素异常变化	共振峰范围及 F2/F1 范围随病情严重程度缩小，语音持续时间延长
整体语音表现	音量稳定均衡，语音连贯流畅，发音清晰易懂，可懂度高	音量不均，语音断裂且停顿不规则，发音模糊，理解困难，交流功能受限

双侧上运动神经元受损所致，临床表现是发声吃力、音调拉长，以及鼻音过重，脑卒中或者脑瘫就属于这种情况；弛缓型则是下运动神经元受损造成的，其特点在于不自然的停顿、气声明显和辅音错误，颅神经疾病就属于此类；共济失调型与小脑受损有关，会出现韵律失常，语调异常，以及发音中断的现象；运动过强型是由基底节病变引起的，亨廷顿病就属于这种情况；运动减弱型在帕金森病患者中较为常见，其主要表现为声音单调、重音减小，混合型则是多种神经受损并存，比如肌萎缩侧索硬化症（Amyotrophic Lateral Sclerosis，ALS），该类型兼具弛缓型与痉挛型的特点。

构音器官存在形态异常时，就会引发器质性构音障碍，其表现为声门破裂音、鼻腔构音，以及咽摩擦音，这种情况在腭裂患者身上比较多见。

构音器官没有明显的形态或者运动方面的异常，听力也是正常的，这些常常是功能性构音障碍的病因特点，其主要表现就是存在一些固定的构音错误，还会有鼻腔构音的情况。这种情况多发生在学龄前儿童身上，不过它的具体病因现在还不是很清楚。

2 构音障碍声学特征

2.1 构音障碍者的发音特征

构音障碍属于由神经系统损伤引发的言语障碍，常常会出现发音含糊不清，语音节奏错乱、音高起伏不规则，以及音量不稳定的状况[13]，这些异常表现造成构音障碍者的发言同正常的发音存在明显差别，在日常生活中的对话过程中遭遇更多阻碍[14]，有关具体声学特征差异的详细分析如表1所示[15–17]，构音障碍者的语音特征往往与其障碍类型及其严重等级紧密相连，发音特征表现出很高的个体特异性[18]。

2.2 特征提取方法的改进方向

特征提取方法的持续改进是提升构音障碍语音识别精度的关键。尽管如梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，

MFCC)、常量子带倒谱系数(Constant Q Cepstral Coefficients，CQCC)这样的传统特征提取方法在一定程度上能够有效地捕捉语音关键信息，但在构音障碍语音识别领域应用还存在着一系列的问题[19]。以下是一些当前特征提取方法的改进方向。

2.2.1 现有特征提取方法的局限性

MFCC的局限性[19]：MFCC虽在语音识别中广泛应用，但处理构音障碍语音时仍存在局限性。在说话人无关的情况下，MFCC的表现不如CQCC，并且它对于构音障碍语音中与严重程度有关的时序细节(像节奏变化之类)的捕捉能力比较薄弱。

CQCC的局限性[19]：与MFCC相比，CQCC具有较高的频谱分辨率，特别在低频、高频部分细节捕捉方面表现更为突出。然而，CQCC在复杂背景噪声中鲁棒性较差，其在嘈杂环境的应用效果会受到影响。

2.2.2 深度学习特征提取

随着深度学习技术的不断发展，传统的特征提取方式渐渐被依靠深度学习的自动化提取方式所替代，这样就能减轻对手工特征提取的依赖，并提高构音障碍语音识别的准确率[13]。端到端(E2E)模型能够直接把原始语音信号映射成文本，这相比传统语音识别而言，简化了很多步骤，比如特征提取、声学建模，以及解码等[20]。这种方法借助深度神经网络学习音频特征与语言模型的联合表现形式，极大地缩减了人工干预的次数，也改进了识别的精度[21]。尤其是在自监督学习的推动下，语音识别技术取得了显著进展。

卷积神经网络(Convolutional Neural Network，CNN)以及长短期记忆网络(Long Short-Term Memory Network，LSTM)这两个模型在端到端模型里比较常见。CNN可以自动从语音的频谱图当中学习语音的局部特征，很适合用于识别构音障碍语音里的短时频谱变化，而LSTM能够捕捉语音中的时间依赖关系，并且适用于具有节奏、音高的波动变化语音的识别[12,13]，将二者融合起来，可以明显改善对构音障碍语音的识别效果，提升模型的稳定性。

Transformer架构是端到端语音识别模型的关键部分，它所具有的自注意力机制在识别带有构音障碍的语音时表现出特有的优势。与CNN更关注局部特征获取、LSTM更侧重于把握长期依赖不同，Transformer可以既获取语音信号里的长期依赖关系，又能获取全局上下文信息，对于构音障碍者由于发音器官功能损伤造成的语音时序结构模糊的情况，Transformer结构会更适合处理[22]。Conformer是对Transformer进行改良后的架构，它把卷积模块合并在自注意力机制之上，如此一来便能够较好地协调局部特征获取以及全局依赖识别这两个方面，从而能够更加精准地捕捉语音信息，在多个构音障碍数据集中，Conformer比传统方法有着明显的性能优化[23]。

Wav2Vec2.0、HuBERT，以及跨语言XLSR模型等自监督学习方法，在构音障碍语音特征获取方面表现出明显的效果。这些方法的核心优势在于可以利用大量未标注的数据执行预训练，该特性使其特别适合于标注成本高且数据较少的构音障碍语音研究领域[24]。

自监督学习框架Wav2Vec2.0通过对比学习任务实现原始音频波形中的语音表示。在构音障碍语音识别任务里，Wav2Vec2.0相对于传统滤波器组特征(Fbank)而言，在词错误率(Word Error Rate，WER)方面有了大幅改进。HuBERT是一种依靠BERT架构的自监

督语音表示学习模型，它经由类似掩码语言模型的预测任务执行训练。研究显示，HuBERT可学习有效声学特征表示，在多种构音障碍语音数据集上表现良好。跨语言XLSR模型借助56,000小时的音频数据完成训练，这些数据包含53种不同的语言，经过这样训练得到的多语言模型，其内部蕴含着更多相似音素的变化，所以更适合应对构音障碍语音存在的问题[24]。

3 构音障碍严重程度评估

3.1 传统人工评估方法及其根本性局限

在过去数十年间，传统的人工评定方法始终被当作临床判断及诊治监测方面的最高标准。但是，伴随对评定精度要求的持续提升，传统人工评定方法本身存在的局限渐渐凸显出来。

流行评定工具在实际应用时面临着很大困难，在临床上常用的一些工具当中，Frenchay构音障碍评定量表第二版（FDA-2）包含反射、呼吸、唇、腭、喉、舌、可懂度，以及影响因素等八个方面，但是它在跨文化交流适应情况与评定者之间的一致性方面并不可观。相关研究表明，在针对构音障碍的听觉感知评定中，67% 的评定者成对比较结果显示，针对同一个感知特征打分时分差不会超出一个等级；同时，当平均评分处于量表中等分值范围时，评定者之间的一致性会表现出较为明显的降低[25]，这种情况可能会直接左右到判断结果是否可靠以及治疗计划能否精准到位。

从神经解剖定位看，梅奥诊所经典分型框架把构音障碍分为痉挛、弛缓、失调、运动减少、运动过多这些类型[4]，这为认识神经病理与言语表现的关系形成了基础，但是临床上做听知觉分类评估的时候会发现，听者常常被声学相似性影响，并没有严格按照事先设定好的病理类型判断[26]，这样便限制了对实际交流障碍准确预测的能力。

构音障碍言语清晰度评定（Assessment of Intelligibility of Dysarthric Speech，AIDS）以功能为导向，目的在于量化词语及句子的可懂程度[27]。不过，该方法的有效性遭到质疑。研究表明，听者在实际判断构音障碍语音时的正确率往往高于传统转录可懂度所得到的分数。

这些问题体现出传统评定方法在方法论层面存在着极大的局限性。其一，主观评定很大程度上依靠评定者自身的认知及经验，所以不同的评定者对于同一项评定所给出的评分结果可能会有很大的波动[28]，这严重影响到评定结果的可靠性；其二，测量工具自身可能成为干扰变量，正字转录和视觉模拟量表就是如此，即便不同清晰度指标间存在中等相关性，但由于量尺不等值，仍然会产生明显的数值差异，从而影响跨研究比较以及临床解读[29]；其三，单个维度往往无法包含疾病的异质性特征，在帕金森病当中，言语清晰度与疾病严重程度、病程或者运动表型之间就没有稳定的关联[30]。

这些根本存在的问题与挑战使得研究者开始重新考量传统评价方法是否有效，进而积极探寻更客观、更标准的评价方法，如此一来就为机器学习、深度学习参与到构音障碍评价当中形成了应用层面的基础。

3.2 智能学习评估

构音障碍严重程度的智能评定技术已有范式变迁，由传统机器学习转为深度学习与融合方法，这明显改善了评定的客观性与精准度，支持向量机（Support Vector Machine，SVM）属于典型的传统方法，在严重程度分级

中有重要意义。Hernandez等给出一种依靠韵律的自动化方法评定构音障碍严重程度，从英语和韩语数据集中获取音高、语速，以及节奏等韵律特征，并融合随机森林（Random Forest，RF）、支持向量机和神经网络分类器实施自动化评定，相比于以MFCC为基础的基线方法，在英语和韩语数据集上分别达成了18.13%和11.22%的准确率优化，在轻度构音障碍识别方面表现更为优异[31]；Javanmardi等把SVM同wav2vec 2.0特征结合，并应用到UASpeech数据库当中，可以明显改善分类性能[32]；在Nemours语音数据库中，Al-Qatab等比较了六种机器学习分类算法，最高准确率能够达到95.80%[33]。

深度学习不断推进，卷积神经网络经由端到端特征学习冲破传统特征工程的瓶颈，在UASpeech数据集方面，Joshy等设计并提出压缩激发网络，并整合SE模块来自适应校准特征，这种情况下大约可达成10%的性能改善[34]；对于同一数据集，Gupta等采用ResNet处理短时语音片段，得到了98.90%的准确率[35]；在循环神经网络范畴当中，Shih等所提出的CNN-GRU混合模型有着98.38%的准确率，比单个架构（诸如CNN、LSTM）更优[36]，凭借Whisper大规模弱监督预训练编码器的迁移学习手段，Rathod等在UASpeech上也得到了98.02%的准确率，这超过了传统声学特征范例[8]。

智能学习评定技术的产生有效地解决了传统评定方法本身存在的某些固有问题，主观评分者之间存在感知评定差异，测量工具会形成干扰变量，而且在对待疾病异质性方面也存在不足。该技术的主要超越之处在于实现了构音障碍评定的客观化，创建起评定用的智能学习模型，从而进一步改良了诊断的准确率。它重新塑造了构音障碍的评定维度，还为更为精确的构音障碍诊疗构筑了方法论根基。

4 构音障碍语音识别

4.1 构音障碍语音识别定义

构音障碍语音识别（Dysarthric Speech Recognition，DSR）希望把构音障碍者的非标准语音自动转录成文本，这对构音障碍而言非常关键。与标准发音相比，构音障碍语音在声学-语音学层面表现出语速不均、发音含糊、节律异常等显著特征，使得现存的通用语音识别系统性能大幅下滑，无法满足实际应用需求，当下有关构音障碍语音识别的研究主要集中在四个方面：① 创建并扩充构音障碍语音数据库；② 针对构音障碍语音，实现数据增强；③ 针对性建模策略（如说话人自适应）；④ 利用多模态或者多流建模等技术针对构音障碍语音实现创新，并提升性能，其具体步骤见图4。

4.2 构音障碍语音数据库

构音障碍语音数据库的创建对于此领域研究的发展十分关键，伴随研究不断推进，技术持续改善，数据库的规模、质量及其丰富程度均得到优化，给诸多研究都提供了重要支持，表2列举了一些常见的构音障碍数据库。

Nemours数据库由Menéndez - Pidal等

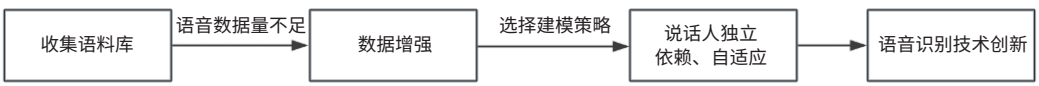

图4 提高构音障碍语音识别准确率具体流程

表 2 构音障碍相关数据库

数据库名称	语言	数据类型	数据规模	特性
Nemours	英语	语音音频	11 名患者；74 条短句 +2 段段落	含音素标记的构音障碍库
UASpeech	英语	音频 + 视频	19 名患者；765 个孤立词	首个视听构音障碍数据库
TORGO	英语	语音 + 发音器官运动数据	7 名患者 + 对照组；共约 23 小时	多模态生理数据采集
CDSD	中文	音频 + 视频	44 名患者；124 小时	最大中文构音障碍库
DEED	英语	多模态情感数据	4 名患者；六种情感 + 中性	首个情感构音障碍库
MSDM	中文	音视频同步	25 名患者 +25 名对照；共 77.1 小时	专注亚急性脑卒中群体
MDSC	中文	音频	21 患者 +25 健康组；共 17 小时	专为语音唤醒设计
VOC-ALS	意大利语	音频	102 患者 +51 健康组；1224 个语音信号	目前最全面的 ALS 语音信号数据库
CUDYS	粤语	词汇 / 句子 / 段落	16 名患者 +5 名对照；共约 10 小时	首个粤语构音障碍库
EasyCall	意大利语	语音交互数据	24 名患者 +31 名对照；共约 21386 个音频记录	智能手机交互专用
PC-GITA	西班牙语	音频	50 患者 +50 健康组	帕金森病专项研究

创建，其中包含11位男性构音障碍患者的74条短无意义句子以及2段连续段落，该数据库在词和音素层级实施了标注，不过因为音素分布较为稀疏，所以给训练ASR模型以及相关研究带来了一定的困难[37]。

UASpeech数据库①由伊利诺伊大学开发，其中收录了19位脑瘫构音障碍患者的语音数据，共计765个孤立词，这些词汇涵盖类型多种多样，用以提升音素序列的丰富性，此数据库为首个集面部视频与语音音频的视听语音数据库，在视听融合的构音障碍语音识别任务中有全面的应用，基于清晰度评估得出的准确率，被试被划分为四个层级：非常低（Very Low）、低（Low）、中等（Mid），以及高（High）[38]。

TORGO数据库②由多伦多大学有关部门与医院共同创建，其中涉及7位患脑瘫或者肌萎缩侧索硬化症从而产生构音障碍的人员，该数据库包含丰富的语音数据，而且还有经由诸多设备收集到的发音器官运动相关数据等，其语音素材包含多种类型，可被用于维持基线水平的能力，执行声学方面的研究等诸多用途[39]。

CDSD③（Chinese Dysarthric Speech Database）属于中文构音障碍语音数据库，由Wang等所创建，其中涵盖了44位患者共计124个小时的音频以及9个小时的视频数据，此数据库被划分为Part A、Part B两个部分，Part A包含44位说话者，每位说话者对应1个小时的音频，合计为44个小时的音频量，而且还有9位说话者的9个小时同步视频；Part B则是从Part A的44人当中挑选出8人，让他们各自录制10个小时的音频，这样就得到了总共80个小时的音频数据，这个数据库利用专业设

① https://speechtechnology.web.illinois.edu/uaspeech/
② https://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html
③ http://melab.psych.ac.cn/CDSD.html

备、智能手机两种场景采集数据，从而兼顾了数据质量及生态有效性[40]。

DEED④（Dysarthric Expressed Emotional Database）是首个多模态情感构音障碍语音数据库，由Alhinti等所创建，使用英式英语，其中收录了4名患者录制的视听资料，包含了六种基本情感表达（愤怒、厌恶、恐惧、快乐、悲伤、惊讶）以及中性表达，这个数据库给情感语音识别研究带来了新路径[41]。

MSDM数据库⑤（Mandarin Subacute Stroke Dysarthria Multimodal Database）专门针对亚急性脑卒中患者进行数据采集，由Liu等创建，用于收集亚急性脑卒中患者的资料，这个数据库收录了25位亚急性脑卒中构音障碍者以及25位健康人的音频-视频同步资料，总时长达77.1小时，还包含了人口统计学特征、认知评定、临床言语评定，以及病灶部位等信息，该数据库与众不同之处在于，其关注的是脑卒中的某个特定时期内的患者人群，并且标注的信息比较全面，所以给脑卒中后的语音复健研究赋予了可靠的来源[42]。

MDSC数据库⑥（Mandarin Dysarthria Speech Corpus）由Gao等创建，此数据库针对语音唤醒、控制命令及其相关语料的采集展开，特别收录了10个唤醒词、355个非唤醒词，该数据库含有21位构音障碍说话者共计9.4小时的录音，还有25位健康对照组合计7.6小时的录音，这给开发适合构音障碍患者的智能家居语音控制系统赋予了数据支持[43]。

VOC-ALS数据库⑦专门针对肌萎缩性侧索硬化症（Amyotrophic Lateral Sclerosis，ALS）患者，其中纳入了153位参与者，具体而言是102位ALS患者以及51位健康对照者，这个数据库包含了共计1224个语音信号，它通过智能手机应用程序采集这些数据，而且其中涉及持续元音发音以及音节重复之类的任务内容，所以给ALS疾病的进程观测和早期判断提供了关键的数据支撑[44]。

CUDYS语料库⑧（Chinese University Dysarthria Corpus）是首个粤语构音障碍语音数据库，由香港中文大学开发。这个语料库涵盖词汇、句子，以及段落这三种任务类型，给粤语地区构音障碍者的语音识别和康复训练给予了基本数据，再加上外部的粤语正常语音语料库之后，该系统在粤语构音障碍语音识别任务里的表现要比商业语音识别API更好[45]。

近年来涌现的其他重要数据库有：EasyCall语料库[46]（语言为意大利语，专门设计用于智能手机语音控制应用的构音障碍语料库，包含55名说话者）、PC-GITA数据库[47]（语言为西班牙语，重点着眼于帕金森病构音障碍的研究，里面涉及100名参与者）等，这些多语言数据库不断发展，充实了构音障碍语音识别研究的资源，给跨语言语音障碍研究赋予了重要支持。

4.3 构音障碍语音处理中的增强方法

在构音障碍语音研究当中，由于难以取得真实构音障碍的数据，而且样本数量较少，所以模型训练及推理遭遇着很大的挑战。要解决因数据稀缺、信号质量差而产生的性能问题，研究者可以从两个方面展开努力：一方面通过语音数据增强扩充训练样本；另一方面通过语音信号增强改善语音清晰度，从而让模型更好地识别构音障碍语音。下文将阐述这两个方面在构音障碍语音处理过程中的主要技术发展情况。

④ https://sites.google.com/sheffield.ac.uk/deed
⑤ https://huanraozhineng1.github.io/MSDM/
⑥ https://www.aishelltech.com/AISHELL_6B
⑦ https://doi.org/10.7303/syn53009474
⑧ https://www1.se.cuhk.edu.hk/~khwong/dysarthria.html

4.3.1 构音障碍语音数据增强技术

语音数据增强是指在不改变语义的前提下，对语音信号施加变换或扰动，生成具有多样性的新样本，从而提高语音识别模型的泛化能力。因构音障碍数据极为稀缺，因此数据增强在该领域尤为重要。

传统方法包括音速扰动（Speed Perturbation）、频谱掩蔽（SpecAugment）、加入背景噪声、音调变换、时间拉伸等技术。这些方法虽简单，但已经有实验证明，这些方法能在扩充数据量的同时，有效地提升模型的鲁棒性和准确率。

近年来，非平行语音转换技术逐渐用于模拟更多样化的病理语音表达方式，非平行语音转换是指在源说话人与目标说话人没有相同语料的情况下，通过特征解耦等方法进行语音转换。也就是只对音色、音调等信息进行转换，而文本内容不发生改变。例如，MaskCycleGAN-VC 等循环一致性生成对抗网络（CycleGAN）方法，在无配对文本内容样本的情况下实现了健康语音与病理语音之间的转换，从而扩展了训练数据的分布范围[48]。

对抗性数据增强也不失为一种表现优异的语音数据增强方式。Wang 等提出了结合对抗性样本生成与预训练识别模型微调的联合增强策略，在 UASpeech 数据集上实现了目前最低的词错误率（16.53%）[49]，为数据增强方法在构音障碍语音识别中的应用树立了新标杆。

4.3.2 构音障碍语音信号增强技术

语音信号增强致力于提升原始语音的清晰度与可懂度，尤其在实际推理或应用场景中显得尤为关键。由于构音障碍语音具备特殊声学特性，标准语音增强方法往往难以直接适用，需针对性改进。

谱减法改进：Boll 提出的单带谱减法是语音增强的经典方法[50]，后来的多带谱减法通过把语言频谱划分为多个频带，这样就可以针对不同频段的噪声强度进行自适应处理，但这些方法直接应用于构音障碍语音会抑制关键语音特征。Chadha 等提出在传统的多带谱减法中优化关键参数（即谱底参数，防止把有用的语音信息当成噪声，这在构音障碍者中更为常见，即被当作噪声过滤掉），对于构音障碍语音能够达到更好地语音信号增强效果[51]。

自适应维纳滤波：传统维纳滤波在处理含噪声的构音障碍语音时，很容易在一定程度上破坏其中的辅音，使其失真。Park 等提出基于辅音-元音（CV）分类的自适应滤波方法，通过结合语音活动检测与元音起始点估计这两种方法，实现了更细粒度的滤波策略，在不同信噪比条件下，准确率均有所提升[52]。

基于发音器官建模：Rudzicz 提出了在原来动态贝叶斯网络框架下引入发音知识，将声学信号与发音器官运动建模结合，提升了构音障碍语音识别的准确率[53]。

4.3.3 深度学习推动下的增强技术融合

深度学习方法作为现在研究的热点领域，在上述两个增强方向都取得了显著进展。

CNN 增强模型：Wang 等提出的深度卷积网络能够学习从构音障碍频谱图到清晰语音频谱图的非线性映射，在语音识别准确率方面提升超过 10%[54]。

Transformer 专用架构：Shahamiri 等设计了定制化 Transformer 架构，通过结合迁移学习技术与参数冻结策略，在数据量极其有限的条件下仍实现了识别性能地极大提升，证实了数据增强与模型结构联合优化的重要性[55]。

4.4 说话人建模策略

在构音障碍语音识别系统设计中，说话人建模策略至关重要，常见方法可分为说话人独立、说话人依赖、说话人自适应，以及基于严重程度分层的混合策略，如图5所示。

4.4.1 说话人独立与说话人依赖模型

说话人独立（Speaker-Independent，SI）方法试图构建一个通用模型处理所有用户的语音，此方法在健康语音的语音识别方面已取得了很大成果。但在构音障碍语音识别时，这种方法仍面临着巨大的挑战，构音障碍语音存在极高的个体差异性，这令传统群体建模方法很难取得预期效果，每个患者的发音缺陷模式、补偿策略等都不相同，使得模型的泛化能力大幅降低。

说话人依赖（Speaker-Dependent，SD）方法通过为每个用户构建专门的模型解决个体差异问题。此方法在构音障碍语音识别领域颇具优势，可以有效地获取患者独有的语音表达模式及发音障碍特征。商用说话人独立ASR系统（如Siri、Google Assistant、Alexa）在处理构音障碍语音时表现不佳，而说话人依赖系统（如mPASS平台）通过用户定制声学模型通常可获得显著更好的性能[56]。

4.4.2 自适应建模：经典方法与预训练–微调

说话人自适应（Speaker-Adaptive，SA）提供了介于SI与SD之间的折中方案。传统上，SA指在通用模型的基础上，利用最大似然线性回归或最大后验等技术，使用少量目标说话人的数据作快速参数调整，从而缓解说话人独立模型很难解决的个体差异性的问题，同时避免说话人依赖方法对大量个人数据的依赖[57]。

近年来，随着自监督预训练模型的普及，“预训练–微调”（Pre-Training & Fine-Tuning，PT-FT）已成为主流自适应新范式。该方法首先在大规模健康或混合语音上训练通用表征（如wav2vec 2.0、HuBERT），随后用极少量构音障碍语音进行轻量化微调或全参数微调，即可在学习通用特征的技术上快速适

图5 说话人建模策略

配不同说话人。实验表明，PT-FT 在数据效率、鲁棒性和性能上限方面均优于传统SA，且可与分层建模策略无缝结合——先按构音障碍严重程度进行群体微调，再视数据规模大小进一步个体化[58]。

4.4.3基于严重程度的分层建模

将患者按病症严重程度或具体障碍类型分组，针对不同层级群体预训练或微调模型，可兼顾群体数据量及相似性，减少每个模型对个体数据量的依赖[59]。

4.5 构音障碍语音识别模型技术演进

相较于将通用语音识别模型直接用于微调构音障碍语音的策略，专门面向构音障碍者的技术创新从特征选择与提取、建模结构到训练机制方面均进行了深度定制，以更好地应对构音障碍者语音的高度异质性及异常的发音模式。

4.5.1多模态融合技术

多模态融合技术利用视觉、生理信号等多种信息提升构音障碍语音识别的性能。视听融合是常见的多模态方法，通过结合唇部运动与声学信息提升识别准确性。更先进的方法还考虑了发音器官的运动信息、肌电信号等生理特征，构音障碍可结合的信息如图6所示。多阶段音视频融合方法结合预训练模型，在处理构音障碍语音时显示出强大的性能[60]。

图 6 多模态信息融合

多阶段音视频融合（MAV-HuBERT）技术把视觉特征扩展至整个面部功能区，并非仅仅局限于唇部运动。该方法结合预训练AV-HuBERT模型，在中等程度构音障碍语音中实现了明显的WER降低[60]。

4.5.2多流声学建模

Yue等开发的多流声学模型通过倒谱提升实现声源与声道成分的分离，利用原始幅度谱的声源和声道成分进行分离-融合建模。该方法相比MFCC基线实现了明显的WER改善，同时有效归一化了构音障碍语音中的说话人属性变异[21]。

SepFormer-SEGAN集成技术结合基于Transformer的语音分离与对抗学习进行构音障碍语音增强，实现了优异的增强效果，为后续识别提供了高质量的输入[61]。

4.5.3跨语言自监督学习

针对构音障碍语音数据稀缺问题，跨语言自监督表示（XLSR）模型通过在多种语言上的预训练，学习语言无关的语音特征模式。Hernandez等的研究表明，该方法在英语、西班牙语和意大利语构音障碍数据集上都实现了显著的WER降低[24]。Hu等发展了输入特征融合技术，通过声学前端与自监督学习（SSL）表示的帧级联合解码，借助领域微调的自监督模型及其特征，实现了显著的绝对WER降低和相对改进[58]。

4.5.4个性化适应策略

少样本学习方法在构音障碍语音个性化中取得重要突破。Google的研究表明，通过选择性层微调，肌萎缩侧索硬化症的患者可实现

显著的相对WER改进，仅使用少量训练数据即可达到明显的改进效果[62]。

对抗性数据增强技术通过预训练ASR系统的微调实现快速适应，而基于光谱基GAN的非平行数据增强策略通过分解语音特征提供了有效的跨患者泛化能力[49]。

这些专门化创新将构音障碍语音识别的性能推向新的高度，在UASpeech等语料库上实现了较高的准确率，即使对于极低可懂度患者也能维持相对合理的WER水平，为临床应用提供了切实可行的技术基础[61]。

5 构音障碍语音修复与语音合成

5.1 构音障碍语音修复

语音修复一般指针对语音增强或语音合成以及语音识别前处理中的任务，以提高语音质量或语音的可理解度，使带有噪声干扰、信息缺失、病态语音条件下的声音更清晰、更自然。构音障碍语音修复是指在说话人具有神经肌肉系统伤病或受损而导致其在进行语音表达中呈现发音缺失、代替、扭曲、错位、断联或节奏不稳（即构音障碍问题）之时，对其语音信号运用算法手段将其进行“恢复”或“重塑”，以使得它们接近正常语音而具有较高的可懂度和谈话效率的过程。

语音内容修复转换技术是构音障碍语音修复的关键所在，其技术发展经历了三个阶段的跨越：从2000年至2015年属于统计参数合成时期，这个阶段以隐马尔可夫模型、高斯混合模型为核心，由Yamagishi团队创建的“语音银行”系统利用模型自适应实现早期的个性化合成[63]，虽然这种方法在计算效率以及特征可控性方面存在优势，但因为过度平滑的原因，合成出来的语音韵律较为单一、自然度欠佳。深度学习革命期（2010—2020年）深度神经网络的引入彻底改变了构音障碍语音修复的技术景观。这一时期的标志性进展包括序列到序列模型和注意力机制的应用。Fu等在2017年提出的基于联合字典学习的非负矩阵分解（JD-NMF）算法代表了早期将深度学习技术应用于构音障碍患者的语音转换任务[64]。该方法在训练数据有限的情况下表现出良好的语音可懂度提升效果，为后续深度学习模型的发展奠定了应用基础。2020年至今为基础模型与先进架构时代，此阶段涌现出更多创新方法。基于音素后验图（PPGs）的语音转换是一个重要进展，Chen等提出的深度语音转换系统采用门控卷积神经网络结合PPGs特征，通过WaveRNN声码器合成转换语音，在安静条件下Google ASR识别率平均为81%，参数数量仅为BLSTM模型的35% [65]。跨模态知识蒸馏框架通过文本和语音联合训练实现更稳定转换效果，而预训练语音基础模型的引入实现更好的内容修复效果。Wang等提出利用HuBERT强大的领域自适应能力以提高训练效率，并利用语音单元在离散语言空间中约束构音障碍内容的恢复，显著提升了内容恢复效果[66]。

说话人身份保持技术解决了语音修复过程中保持个人特征的难题，特别是在缺乏患者正常语音的情况下。受大语言模型及零样本语音合成启发，Chen等提出利用神经编解码器语言模型改善语音重建，有效提升了说话人相似度和韵律自然度[67]。基于对比学习说话人嵌入技术的最新进展，Keshvari等的方法使用对比学习提取说话人嵌入，采用XLSR表征替代传统滤波器组特征，在中度及中重度构音障碍语音上都实现了显著的MOS评分改善及词错误率降低[68]。

5.2 构音障碍语音合成

语音合成（Speech Synthesis）指的是通过计算机算法把文本或者其他语言表现形式（比如语音）转变成可听语音信号这样一个过程。近年来，语音合成全面用在了语音交互、辅助交流等众多领域，在医学语音处理方面所具有的应用价值格外突出。在构音障碍这个范畴当中，数据稀少是十分明显的一个问题，需要利用语音合成技术扩充构音障碍语音，以满足识别、评价，以及诊疗相关任务所需的数据量，进而从某种意义上减轻了对于大量标注数据的依赖程度。

构音障碍语音合成领域面临诸多技术难题，关键问题在于构音障碍语音声学特征丰富，而语音数据却较为稀缺，脑瘫、帕金森病、肌萎缩侧索硬化症等病症所造成的病理差异十分个体化，它们的声学特征存在很大差别，所以模型要有很强的建模能力以产生多种构音障碍声学特征。但当前可用的构音障碍语音资源无论是数量还是种类都远远少于健康语音语料库。近些年来，生成式人工智能促使这个领域快速发展。Wang等借助概率扩散模型达成从正常语音向构音障碍语音的转变[69]，该模型保留说话者身份并体现严重程度，用作数据加强手段以改善识别效果。Leung等采用扩散模型（Grad - TTS等）合成构音障碍语音实施数据加强，有效地对Whisper等大型ASR模型执行了微调[70]。Zhang等提出依靠长距离非平稳变分自编码器（LNVAE）的模型，经由编码长时间音素依赖关系和潜变量噪声干扰，获取构音障碍语音的非平稳波动情况[71]。

5.3 修复或合成后语音质量评估

构音障碍经语音修复或者合成之后，一般要对语音质量进行评估，以确定是否达到预期目标，评估过程需综合考量语音质量、可懂度、自然度等诸多方面。客观评估通过标准化指标量化改善效果，主观评估则是最终判定标准，需要结合专业听者及普通听者的评估结果，具体评价指标如表3所示[72]。

表 3 修复或合成语音评估类型与指标

评估类型	评估方法	评估方式	应用场景及注意事项
主观评价方法	MOS（平均意见分）	被试听众对修复或合成语音主观打分	需多样化样本，用于初步评估语音整体可接受度
	CMOS（比较平均意见分）	对两个系统句子两两对比、七分制打分	对质量差异更敏感，用于不同语音合成系统对比
	ABX Test	参与者三次听觉对比中选 A/B 与 X 匹配	评估音频编解码器、语音合成系统效果
	MUSHRA（多刺激隐藏参考和锚点法）	评估者对参考音频与多个样本评分，寻找最接近参考的样本	对比多个音频样本与参考音频的相似度
客观评价方法	MCD（梅尔倒谱畸变）	计算合成与天然 mel 倒谱序列的 MCD，值越小自然性越高	需结合其他指标使用
	PESQ（语音质量感知评估）	计算原始与处理后语音差异得质量分数	衡量语音编解码器、通信系统性能
	STOI（短时客观可懂度）	比较原始与失真 / 噪声语音频谱相关性，评估可懂度	用于语音合成可懂度、识别率评估，以及语音增强、降噪等场景
	LLR（对数似然比）	评估模型生成语音是否属于给定语音分布	判断生成语音与特定语音分布匹配度

6 挑战与展望

智能语音技术近年来虽发展快速，给构音障碍领域带来新前景，但依然无法完全达到实际应用标准，还受数据稀缺、个性化适配缺乏等诸多技术难题限制，日后研究可从如下几方面达成研究超越与进程。

构音障碍相关语音技术面临的关键难点在于，高质量训练数据极为稀少，且难以弥补个体差异。构音障碍语音标注数据的数量远小于训练正规语言模型所需的数据量，其质量标准也缺乏统一准则。构音障碍人群在发音机制、语速节奏、音调变化等方面存在明显差别，所以很难训练出适合全部构音障碍者的模型。应对这种困境，既要不断搜集数据，也要深入探究少样本学习、迁移学习等技术，以解决因数据量不够、个体差异造成的模型泛化问题，还可以通过自适应特征获取，动态模型微调等手段达成“一人一模型”的精确适配，大幅优化语音识别的准确率。

构音障碍者的语音从声学层面看与健康对照组存在很大差别，比如能量分布不均、音节时长不规则等情况，所以当前通用的识别模型用来识别这类语音的时候，性能就会大幅下降。现在语音识别技术被更多地用在语音唤醒、指令发出等方面，但是构音障碍者常常无法很好地使用这些功能，这样就有可能加大数字鸿沟，所以要尽早设计出适合构音障碍者以及其他特殊群体的智能语音识别系统。

构音障碍相关辅助器具的硬件、软件设计要充分顾及使用者的特别需求。硬件方面应具有轻便、低能耗、高耐用这些特性，使得设备可随身携带且能够长时间使用，也要把成本控制在合理范围内。软件方面，所设计的交互界面应具有易操作性、高容错率等特性，通过精简操作流程，提升视觉反馈等途径来减小操作难度。硬件与软件相互结合，以保障设备在复杂声学环境中具有稳定的性能。

人机交互的个性化表达辅助系统是重要的发展方向，把认知行为治疗理论与人工智能技术结合，可以形成具有情感计算能力的智能陪伴系统，为构音障碍者提供心理疏导服务。这种系统不但能够识别使用者的情绪状况，给予及时的心理支撑，而且可以通过对话交流增加语言训练的次数，逐步改善其发音清晰度，提高表达自信心，助力其融入社会。

构音障碍方向需要形成统一的技术规范及性能评价体系，为后续的数据采集及模型应用给予指引。在设计语言系统时，要考虑不同类型的构音障碍群体的需求，避免某个群体在数字浪潮中被边缘化。此外，还应通过政策引导、社会宣传等方式优化全社会对构音障碍群体的包容程度。

7 结论

智能语音技术在构音障碍领域不断发展，这会在很大程度上帮助构音障碍者更有效地融入人际交流之中，从而改善其生活质量以及参与社会的程度。但是，当前其仍然遭遇不少瓶颈，比如数据量不足且个体间差异化较大、算法不够稳定、缺乏临床适应性等情况。如果通过划定构音障碍的疾病种类，剖析其声学特征并考量严重程度，就能为分类研究及定向分析带来一定的帮助。加强构音障碍语音的识别率是推进这个领域发展的关键目标之一。因此，要有更大规模、更高质量的构音障碍语音数据，还要遵照说话人数目、数据量以挑选更恰当的建模方案，也要对语音识别模型执行进一步的改良。构音障碍语音修复通过去除噪音、纠正

异常发音，可以提升患者语音的清晰度与可理解度，从而改良实际交流的体验。构音障碍语音合成可用于生成多种病理语音样本，用以填补真实数据的短缺，为数据增强及模型训练提供支撑。将这两者结合，既有益于冲破当前的技术难题，又为设计针对构音障碍群体的个性化辅助交流系统形成基础。日后，伴随大量高质量数据资源持续采集、个性化建模方法逐渐成熟，智能语音技术会给构音障碍群体带来效率更高、通用性更强、更具社会包容性的技术支持。

参考文献

[1] DUFFY J R. Motor speech disorders: substrates, differential diagnosis, and management[M]. St. Louis, Mo: Mosby, 2013.
[2] KIM Y, KIM M, KIM J, et al. Smartphone-based speech therapy for poststroke dysarthria: pilot randomized controlled trial evaluating efficacy and feasibility[J]. Journal of Medical Internet Research, 2024, 26: e56417.
[3] MITCHELL C, BOWEN A, TYSON S, et al. Interventions for dysarthria due to stroke and other adult-acquired, non-progressive brain injury[J]. The Cochrane Database of Systematic Reviews, 2017, 1(1): CD002088.
[4] DARLEY F L, ARONSON A E, BROWN J R. Differential diagnostic patterns of dysarthria[J]. Journal of Speech and Hearing Research, 1969, 12(2): 246-269.
[5] LIU H M, TSAO F M, KUHL P K. The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy[J]. The Journal of the Acoustical Society of America, 2005, 117(6): 3879-3889.
[6] LANSFORD K L, LISS J M. Vowel acoustics in dysarthria: mapping to perception[J]. Journal of Speech, Language, and Hearing Research, 2014, 57(1): 68-80.
[7] ENDERBY P, PALMER R. FDA-2: frenchay dysarthria assessment-second edition[M]. PRO-ED, 2008.
[8] RATHOD S, CHAROLA M, VORA A, et al. Whisper features for dysarthric severity-level classification[C]// INTERSPEECH 2023. ISCA, 2023: 1523-1527.
[9] QIAN Z P, XIAO K J. A survey of automatic speech recognition for dysarthric speech[J]. Electronics, 2023, 12(20): 4278.
[10] POPOV V, VOVK I, GOGORYAN V, et al. Grad-TTS: a diffusion probabilistic model for text-to-speech[C]// Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021.
[11] YORKSTON K M. Management of motor speech disorders in children and adults[M]. Austin, Tex: Pro Ed, 1999.
[12] ENDERBY P. Chapter 22 - Disorders of communication: dysarthria[M]//Handbook of Clinical Neurology. Elsevier, 2013: 273-281.
[13] KENT R D, WEISMER G, KENT J F, et al. Acoustic studies of dysarthric speech: methods, progress, and potential[J]. Journal of Communication Disorders, 1999, 32(3): 141-186.
[14] ROWE H P, GUTZ S E, MAFFEI M F, et al. Characterizing dysarthria diversity for automatic speech recognition: a tutorial from the clinical perspective[J]. Frontiers in Computer Science, 2022, 4: 770210.
[15] YILMAZ E, MITRA V, BARTELS C, et al. Articulatory features for ASR of pathological speech[C]//Interspeech 2018. ISCA, 2018: 2958-2962.
[16] THOPPIL M G, KUMAR C S, KUMAR A, et al. Speech signal analysis and pattern recognition in diagnosis of dysarthria[J]. Annals of Indian Academy of Neurology, 2017, 20(4): 352-357.
[17] LANSFORD K L, LISS J M. Vowel acoustics in dysarthria: speech disorder diagnosis and classification[J]. Journal of Speech, Language, and Hearing Research, 2014, 57(1): 57-67.
[18] BRONAGH BLANEY J W. Acoustic variability in dysarthria and computer speech recognition[J]. Clinical Linguistics & Phonetics, 2000, 14(4): 307-327.
[19] JOSHY A A, RAJAN R. Automated dysarthria severity classification: a study on acoustic features and deep learning techniques[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 1147-1157.
[20] TAKASHIMA R, SAWA Y Y, AIHARA R, et al. Dysarthric speech recognition using pseudo-labeling, self-supervised feature learning, and a joint multi-task learning approach[J]. IEEE Access, 2024, 12: 36990-36999.
[21] YUE Z J, LOWEIMI E, CHRISTENSEN H, et al. Acoustic modelling from raw source and filter components for dysarthric speech recognition[J]. ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2968-2980.
[22] IRSHAD U, MAHUM R, GANIYU I, et al. UTran-DSR: a novel transformer-based model using feature enhancement for dysarthric speech recognition[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2024, 2024(1): 54.
[23] WANG Q L, ZHONG Z H, SINGH S, et al. Dysarthric speech conformer: adaptation for sequence-to-sequence dysarthric speech recognition[C]//2025 IEEE International Conference on Acoustics, Speech and Signal Processing, April 6-11, 2025, Hyderabad, India. IEEE, 2025.

[24] HERNANDEZ A, PÉREZ-TORO P A, NOETH E, et al. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition[C]//Interspeech 2022. ISCA, 2022: 51-55.
[25] BUNTON K, KENT R D, DUFFY J R, et al. Listener agreement for auditory-perceptual ratings of dysarthria[J]. Journal of Speech, Language, and Hearing Research, 2007, 50(6): 1481-1495.
[26] LANSFORD K L, BERISHA V, UTIANSKI R L. Modeling listener perception of speaker similarity in dysarthria[J]. The Journal of the Acoustical Society of America, 2016, 139(6): EL209.
[27] NARENDRA N, ALKU P. Automatic intelligibility assessment of dysarthric speech using glottal parameters[J]. Speech Communication, 2020, 123: 1-9.
[28] BORRIE S A, MCAULIFFE M J, LISS J M. Perceptual learning of dysarthric speech: a review of experimental studies[J]. Journal of Speech, Language, and Hearing Research, 2012, 55(1): 290-305.
[29] STIPANCIC K L, TJADEN K, WILDING G. Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls[J]. Journal of Speech, Language, and Hearing Research, 2016, 59(2): 230-238.
[30] MILLER N, ALLCOCK L, JONES D, et al. Prevalence and pattern of perceived intelligibility changes in Parkinson's disease[J]. Journal of Neurology, Neurosurgery, and Psychiatry, 2007, 78(11): 1188-1190.
[31] HERNANDEZ A, KIM S, CHUNG M. Prosody-based measures for automatic severity assessment of dysarthric speech[J]. Applied Sciences, 2020, 10(19): 6999.
[32] JAVANMARDI F, TIRRONEN S, KODALI M, et al. Wav2vec-based detection and severity level classification of dysarthria from speech[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing, June 4-10, 2023, Rhodes Island, Greece. IEEE, 2023: 1-5.
[33] ALI AL-QATAB B, MUSTAFA M B. Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features[J]. IEEE Access, 2021, 9: 18183-18194.
[34] JOSHY A A, RAJAN R. Dysarthria severity assessment using squeeze-and-excitation networks[J]. Biomedical Signal Processing and Control, 2023, 82: 104606.
[35] GUPTA S, PATIL A T, PUROHIT M, et al. Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments[J]. Neural Networks, 2021, 139: 105-117.
[36] SHIH D H, LIAO C H, WU T W, et al. Dysarthria speech detection using convolutional neural networks with gated recurrent unit[J]. Healthcare, 2022, 10(10): 1956.
[37] MENÉNDEZ-PIDAL X, POLIKOFF J B, PETERS S M, et al. The Nemours database of dysarthric speech[C]//4th International Conference on Spoken Language Processing. ISCA, 1996: 1962-1965.
[38] KIM H, HASEGAWA-JOHNSON M, PERLMAN A, et al. Dysarthric speech database for universal access research[C]//Interspeech 2008. ISCA, 2008: 1741-1744.
[39] RUDZICZ F, NAMASIVAYAM A K, WOLFF T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria[J]. Language Resources and Evaluation, 2012, 46(4): 523-541.
[40] WAN Y, SUN M Y, KANG X C, et al. CDSD: Chinese dysarthria speech database[C]//Interspeech 2024. ISCA, 2024: 4109-4113.
[41] ALHINTI L, CUNNINGHAM S, CHRISTENSEN H. The dysarthric expressed emotional database (DEED): an audio-visual database in British English[J]. PLoS One, 2023, 18(8): e0287971.
[42] LIU J, DU X X, LU S J, et al. Audio-video database from subacute stroke patients for dysarthric speech intelligence assessment and preliminary analysis[J]. Biomedical Signal Processing and Control, 2023, 79: 104161.
[43] GAO M, CHEN H, DU J, et al. Enhancing voice wake-up for dysarthria: mandarin dysarthria speech corpus release and customized system design[C]//Interspeech 2024. ISCA, 2024: 2465-2469.
[44] DUBBIOSO R, SPISTO M, VERDE L, et al. Voice signals database of ALS patients with different dysarthria severity and healthy controls[J]. Scientific Data, 2024, 11(1): 800.
[45] WONG K H, YEUNG Y T, CHAN E H Y, et al. Development of a Cantonese dysarthric speech corpus[C]//Interspeech 2015. ISCA, 2015: 329-333.
[46] TURRISI R, BRACCIA A, EMANUELE M, et al. EasyCall corpus: a dysarthric speech dataset[C]//Interspeech 2021. ISCA, 2021: 41-45.
[47] OROZCO-ARROYAVE J R, ARIAS-LONDOÑO J D, VARGAS-BONILLA J F. New spanish speech corpus database for the analysis of people suffering from parkinson's disease[C]//Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland. European Language Resources Association, 2014: 342–347.
[48] KANEKO T, KAMEOKA H, TANAKA K, et al. Maskcyclegan-VC: learning non-parallel voice conversion with filling in frames[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing, June 6-11, 2021,Toronto, ON, Canada. IEEE, 2021: 5919-5923.
[49] WANG H M, JIN Z R, GENG M Z, et al. Enhancing pre-trained ASR system fine-tuning for dysarthric speech recognition using adversarial data augmentation[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing, April 14-19, 2024, Seoul, Korea,

Republic of. IEEE, 2024: 12311-12315.
[50] BOLL S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.
[51] CHADHA A N, ZAVERI M A, SARVAIYA J N. Optimized spectral floor in multi-band spectral subtraction for dysarthric speech recognition[C]//2017 4th International Conference on Signal Processing and Integrated Networks, February 2-3, 2017, Noida, Delhi-NCR, India. IEEE, 2017: 245-250.
[52] PARK J H, SEONG W K, KIM H K. Preprocessing of dysarthric speech in noise based on CV–dependent Wiener filtering[M]//Proceedings of the paralinguistic information and its integration in spoken dialogue systems workshop. New York, NY: Springer New York, 2011: 41-47.
[53] RUDZICZ F. Articulatory knowledge in the recognition of dysarthric speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 947-960.
[54] WANG S S, TSAO Y, ZHENG W Z, et al. Dysarthric speech enhancement based on convolution neural network[C]//2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, July 11-15, 2022, Glasgow, Scotland, United Kingdom. IEEE, 2022: 60-64.
[55] SHAHAMIRI S R, LAL V, SHAH D. Dysarthric speech transformer: a sequence-to-sequence dysarthric speech recognition system[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 3407-3416.
[56] JADDOH A, LOIZIDES F, RANA O. Interaction between people with dysarthria and speech recognition systems: a review[J]. Assistive Technology, 2023, 35(4): 330-338.
[57] RAMYA T, LILLY CHRISTINA S, VIJAYALAKSHMI P, et al. Analysis on MAP and MLLR based speaker adaptation techniques in speech recognition[C]//2014 International Conference on Circuits, Power and Computing Technologies, March 20-21, 2014, Nagercoil, Tamil Nadu, India. IEEE, 2014: 1753-1758.
[58] HU S J, XIE X R, GENG M Z, et al. Self-supervised ASR models and features for dysarthric and elderly speech recognition[J]. IEEE Transactions on Audio, Speech and Language Processing, 2024, 32: 3561-3575.
[59] KIM M J, YOO J, KIM H. Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models[C]//Interspeech 2013. ISCA, 2013: 3622-3626.
[60] YU C C, SU X S, QIAN Z P. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 1912-1921.
[61] VINOTHA R, HEPSIBA D, VIJAY ANAND L D, et al. Enhancing dysarthric speech recognition through SepFormer and hierarchical attention network models with multistage transfer learning[J]. Scientific Reports, 2024, 14(1): 29455.
[62] SHOR J, EMANUEL D, LANG O, et al. Personalizing ASR for dysarthric and accented speech with limited data[C]//Interspeech 2019. ISCA, 2019: 784-788.
[63] YAMAGISHI J, VEAUX C, KING S, et al. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction[J]. Acoustical Science and Technology, 2012, 33(1): 1-5.
[64] FU S W, LI P C, LAI Y H, et al. Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery[J]. IEEE Transactions on Biomedical Engineering, 2017, 64(11): 2584-2594.
[65] CHEN C Y, ZHENG W Z, WANG S S, et al. Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system[C]//Interspeech 2020. ISCA, 2020: 4686-4690.
[66] WANG Y J, WU X X, WANG D S, et al. UNIT-DSR: dysarthric speech reconstruction system using speech unit normalization[Z]. arXiv preprint arXiv: 2401.14664, 2024.
[67] CHEN X Y, YANG D C, WANG D D, et al. CoLM-DSR: leveraging neural codec language modeling for multi-modal dysarthric speech reconstruction[C]//Interspeech 2024. ISCA, 2024: 4129-4133.
[68] FATEMEH K, RAHIL M T, HASSAN Z. Enhancement of dysarthric speech reconstruction by contrastive learning[Z]. arXiv preprint arXiv: 2410.04092, 2024.
[69] WANG H L, THEBAUD T, VILLALBA J, et al. DuTa-VC: a duration-aware typical-to-atypical voice conversion approach with diffusion probabilistic model[C]//INTERSPEECH 2023. ISCA, 2023: 1548-1552.
[70] LEUNG W Z, CROSS M, RAGNI A, et al. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis[C]//Interspeech 2024. ISCA, 2024: 2494-2498.
[71] ZHANG D P, ZHANG H C, LU W H, et al. Long-range and non-stationary encoding for dysarthric speech data augmentation[J]. IEEE Journal of Selected Topics in Signal Processing, 2025: 1-16.
[72] KUBICHEK R. Mel-cepstral distance measure for objective speech quality assessment[C]//Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada. IEEE, 1993: 125-128.

作者简介

赵欣然　女，江苏科技大学计算机学院硕士研究生在读，中国科学研究院心理研究所实习生。主要研究方向为构音障碍语音识别。
共同第一作者

刘　柏　男，长春大学计算机科学技术学院硕士研究生在读，中国科学研究院心理研究所实习生。主要研究方向为构音障碍语音识别。
共同第一作者

刘小康　男，中国科学院深圳先进技术研究院先进技术集成所博士研究生在读。主要研究方向包括病理性语音处理（构音障碍、口吃检测）、自动语音识别、说话人识别和多模态语音处理。在IEEE/ACM TASLP、ICASSP、INTERSPEECH等国际期刊及会议发表多篇论文，并多次在语音处理竞赛中获奖。

吴锡欣　男，香港中文大学系统工程与工程管理系助理教授。曾任剑桥大学工程系副研究员、香港中文大学何鸿燊海量数据决策分析研究中心研究助理教授。主要研究方向包括语音合成、情感识别、音频鉴伪等。曾获得INTERSPEECH 2020非母语儿童语音自动语音识别比赛第一名、ACII 2022情感人声爆发识别比赛第一名、ACL 2022 DialDoc文本问答比赛第一名。

燕　楠　男，中国科学院深圳先进技术研究院研究员、博士生导师。主要研究领域为语音人工智能，语音信号处理、儿童言语发育障碍脑机制与调控技术、智能化言语康复技术。近五年来，主持国家重点研发计划、国家自然科学基金面上项目等多个课题，作为核心骨干参与国家自然科学基金重点项目、国家重点研发计划变革性技术关键科学问题等重大项目。发表SCI/EI期刊会议论文100余篇。

王甦菁　男，中国科学院心理研究所副研究员、博士生导师。兼任中国计算机学会理事。获2018年第八届吴文俊人工智能科学技术奖一等奖、2023年度北京市科学技术奖自然科学奖二等奖。北京2022年冬残奥会火炬手，被新华社称为“中国版霍金”，第七届全国自强模范。在国内外重要期刊及学术会议发表论文一百余篇。
通信作者email：wangsujing@psych.ac.cn。