大语言模型时代的无声语音接口：全面分类法与系统综述

Kele Xu、Yifan Wang、Ming Feng、Qisheng Xu、Wuyang Chen、Yutao Dou、Cheng Yang、Huaimin Wang

arXiv:2603.11877v1 [eess.AS] 12 Mar 2026

摘要——**人机交互传统上依赖声学通道，这种依赖会带来对环境噪声、隐私限制和生理性言语障碍的系统性脆弱性。无声语音接口（SSIs）作为一种变革性范式应运而生，通过直接从神经-肌肉-发音连续体解码语言意图，绕过声学阶段。本综述对无声语音接口领域进行了高层次综合，从传统的以换能器为中心的分析转向整体性的意图到执行分类法。我们系统评估了四个关键生理截取点上的传感模态：神经振荡、神经肌肉激活、发音运动学（超声/磁测量），以及通过声学或射频传感进行的普适主动探测。至关重要的是，我们分析了当前从启发式信号处理到潜在语义对齐的范式转变。在这一新时代，大语言模型（LLMs）与深度生成架构作为高层语言先验，用以解决生物信号的“信息稀疏性”和非平稳性。通过将碎片化的生理动作映射到结构化语义潜在空间，现代无声语音接口框架首次接近了真实世界部署所需的词错误率可用性阈值。我们进一步考察了无声语音接口从笨重的实验室仪器向集成于消费级可穿戴设备（如耳戴式设备和智能眼镜）的“隐形接口”的转变。最后，我们概述了一条战略路线图，借助自监督基础模型应对“用户依赖性悖论”，并界定“神经安全”的伦理边界，以在日益接口化的世界中保护认知自由。**

索引词—**无声语音接口、发音传感、深度学习、大语言模型、人机交互、辅助技术。**

一、引言

人机交互（HCI）传统上以声学通道为基础。传统自动语音识别（ASR）系统——从经典随机建模发展到当代端到端神经网络架构——均以高保真音频采集为前提 [1], [2]。然而，这种根本性的依赖对空气传播的声音造成了系统性脆弱性：（i）**环境脆弱性：**在非平稳噪声或高混响环境中，信噪比（SNR）严重受损，系统性能会急剧下降 [3]；（ii）**社会与隐私限制：**必须进行可听发声，导致在公共场所或共享空间中发生敏感信息泄露并产生“社会摩擦” [4]；以及（iii）**包容性障碍：**依赖声学的系统本质上排除了患有严重发声障碍的人群，例如因喉切除术导致失声者或患有神经退行性语音障碍者 [5]。

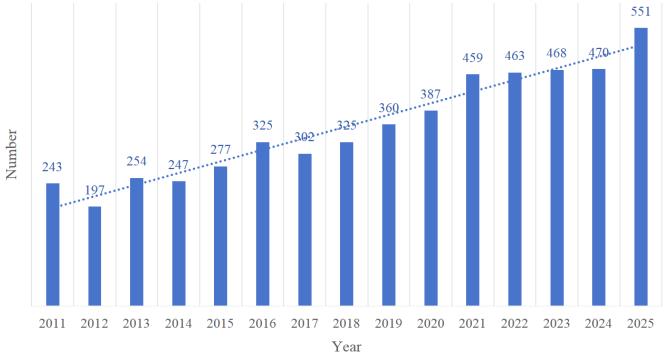


图1：静默语音接口（SSI）领域的论文发表总数。统计数据检索自 Web of Science，检索主题包括“无声语音”“无声语音接口”“SSI”“无声语音识别”“亚发声语音”“无声语音识别”“不可听语音”“发音语音识别”“发音动作到文本”以及“无声语音”，统计时间范围为2011年至2025年。

无声语音接口（SSIs）作为一种变革性范式应运而生，它通过直接从语音产生的神经-肌肉-发音链中解码语言意图，绕过声学通道 [6], [7], [8], [9], [10]。为进一步说明这一领域的快速发展及日益受到的研究关注，如图1所示，2011年至2025年，与无声语音接口相关的年度出版物数量持续增长。语音产生过程始于皮层运动意图，经由传出神经肌肉激活传递，最终形成声道的物理运动学特征（例如舌、嘴唇和软腭）。无声语音接口通过在不同生理阶段截取这一链条——涵盖神经振荡（皮层脑电图/脑电图）、肌电激发（表面肌电）和发音轨迹（超声、磁测量或计算机视觉）——有效地将交际意图与声学

Kele Xu、Yifan Wang、Ming Feng、Qisheng Xu 和 Cheng Yang 任职于国防科技大学并行与分布式计算国家重点实验室、计算机科学与技术学院。他们同时任职于国防科技大学复杂关键软件环境国家重点实验室，地址为中国长沙市，邮编410073。Wuyang Chen 任职于湖南师范大学计算机科学学院，地址为中国长沙。Yutao Dou 任职于湖南大学计算机科学学院，地址为中国长沙。

压力波分离开来。这使得生理前兆能够直接映射为语义或合成的声学表示，从而有效地重新界定语音通信的边界 [9]。

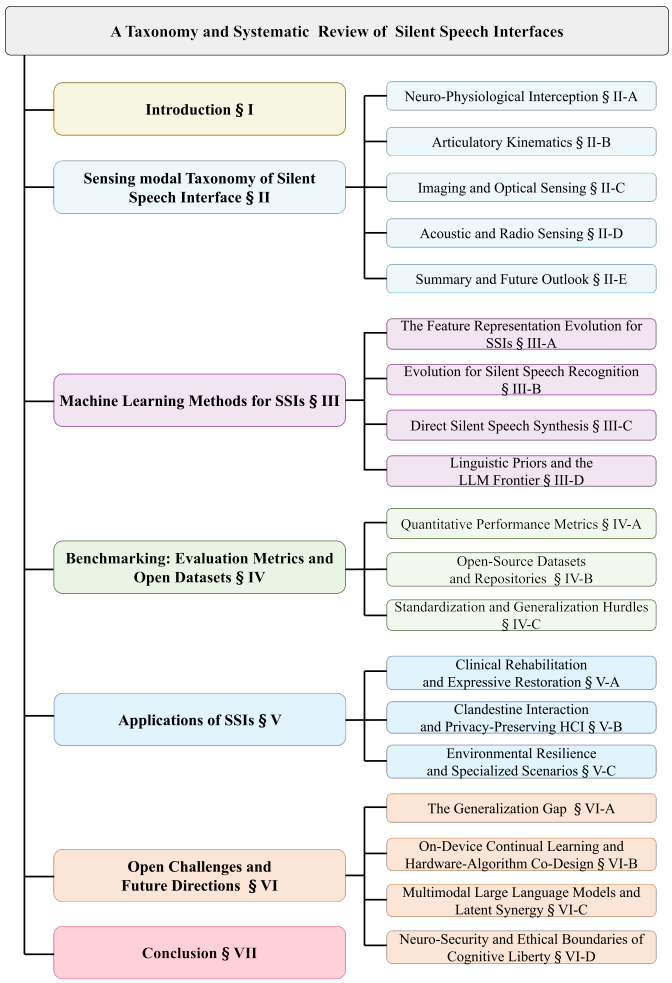


图 2：本文综述的框架。

当前静默语音接口研究的势头由三项相互汇聚的技术与社会需求推动：（i）**高性能神经假体：**对于言语丧失严重的个体而言，无声语音接口是重新获得交流身份的关键桥梁。近期高密度神经解码领域的突破表明，可以从皮层活动中合成自然的语音，其词错误率（WER）正日益接近人类水平性能 [11], [12]。（ii）**保护隐私的普及性：**泛在人工智能的普及加剧了对隐蔽交互模态的需求。无声语音接口支持“无声”指令，使用户能够在完全静默的情况下与数字智能体交互，从而降低敏感环境中遭受声学窃听的风险 [13]。（iii）**声学免疫性与鲁棒性：**通过在非声学域中运行，无声语音接口为“鸡尾酒会问题”提供了内在的解决方案。表面肌电（sEMG）和主动声学感测（AAS）等模态无论环境声压级如何，都能保持信号完整性，从而确保在极端工业或航空环境中的确定性性能 [14], [15]。

过去十年间，无声语音接口的技术格局经历了深刻蜕变，从笨重且局限于实验室的仪器，转向低调、不显眼的消费级可穿戴设备。早期研究主要聚焦于精度高但限制性较强的模态，例如超声舌成像（UTI） [16]或永久磁发音动作分析（PMA） [17]。然而，主动感测和柔性电子学领域的近期创新，推动了面向消费者的形态转变。当前最先进的无声语音接口利用集成于标准耳塞式耳机[14],、智能眼镜 [15]，和腕戴设备 [18] 中的换能器，通过声学回波或惯性测量单元（IMUs）追踪发音动力学。值得注意的是，非接触式磁测量和生物阻抗感测的出现，引入了高灵敏度的替代方案，规避了传统表面肌电（sEMG）固有的皮肤耦合和运动伪影挑战。

随着硬件的发展，大语言模型（LLMs）与深度生成架构的融合也彻底改变了无声语音接口计算后端。长期以来，无声语音接口识别面临的一大瓶颈是“无声隆巴德效应”——缺乏听觉反馈时发音动作中出现的系统性生理变化 [19]。当代框架通过多模态对比学习和大语言模型条件重评分来应对这一问题。通过在统一的潜在空间中，将零散的生理特征与高层语言先验对齐，这些系统能够可靠地重建缺失的发音信息，在开放词汇的连续语音任务中实现前所未有的准确度 [20], [21]。

尽管该领域取得了迅速进展，但仍缺少一项全面的综述，以弥合异构传感模态与最新生成式人工智能范式之间的鸿沟。本文对无声语音接口领域的最新进展进行了系统综述。我们基于言语产生链建立了技术分类法，并对从传统分类方法向现代基于 Transformer 和基于扩散的架构的转变进行了批判性评估。此外，我们还探讨了无声语音接口中新兴的交互范式 [22], ，同时关注信号非平稳性、跨受试者域适应，以及“思维转语音”神经技术的伦理影响等未决挑战。

本综述的具体贡献如下（如图2所示）：

- **统一生理分类法：**我们基于静默语音接口模态在神经-肌肉-发音链中的特定截取点，对其进行严格分类，并对其物理约束与信息容量进行比较分析。

- **算法演进分析：**我们综合梳理了从启发式特征工程到深度表征学习的转变，并重点关注跨模态潜在对齐、自监督预训练以及大语言模型驱动的解码框架。

- **多层面应用场景：**我们阐述了无声语音接口多样的应用场景，涵盖从针对言语障碍的临床修复性神经假体，到公共空间中的安全通信，以及高噪声工业或航空航天场景。
- **基准测试与开放科学：**我们提供了一套完整的

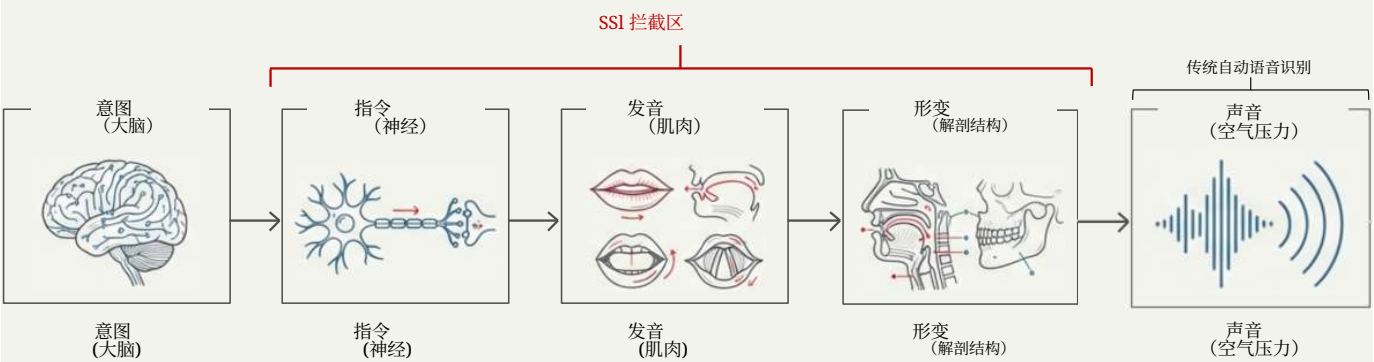


图 3：神经-肌肉-发音连续体与静默语音接口截获边界。该示意图勾勒了语音产生的生理阶段，从皮层意图和神经信号传递，到肌肉激活和解剖结构形变。静默语音接口截获区（红色高亮）界定了相关技术的功能范围：这些技术绕过声学阶段（传统自动语音识别），通过非发声生物信号恢复或增强交流能动性。

经过评估的开源数据集和标准化评估指标资源库，以促进可复现研究和跨模态基准测试。

- **挑战与研究路线图：**我们指出了跨受试者泛化和零样本迁移方面的关键差距，并为下一代稳健且符合伦理的静默语音技术提出了路线图。

II. 静默语音接口感测模态分类

无声语音接口的运行效能本质上取决于其在神经-肌肉-发音连续体上的截获位置 [23]。如图 3 所示，无声语音接口截获区涵盖语音产生的整个非声学轨迹，实际上绕过了声学通道，从而将语言意图与大气压力波分离 [24]。

这些模态构成了一种层级分类体系——如表 I 所示——依据其所换能的物理现象的深度，从近端的皮层振荡到远端的运动学执行 [25], [26]。这一层级揭示了一个根本性的技术矛盾：近端神经解码器拥有更高的自由度（DoF）和语义密度，而远端可穿戴传感器则具有更高的社会接受度 [27]–[28]。从这些生理阶段过渡到功能性接口，需要采用多种换能技术，而每种技术都面临独特的限制，例如“可观测性–普及性悖论”。

A. 神经生理截获

神经生理感测在生物源头捕捉语音意图，绕过声道的物理约束。这些模态必须应对低信噪比（SNR）以及神经流形中高度的受试者间变异。

1) 皮层意图：皮层感知旨在通过监测大脑活动，解码想象或意图表达的语音，在非侵入式技术的可扩展性与侵入式技术的高保真度之间应对根本性的权衡。头皮记录的脑电图（EEG）因其可及性和较高的时间分辨率而颇具吸引力 [29], [30]；然而，颅腔容积导体会严重衰减空间信息，导致皮层表征模糊。因此，近期研究转向神经—语言对齐 [12]，通过自监督预训练将噪声神经嵌入映射到结构化的语言潜在空间，并利用语义先验对解码进行正则化。

相比之下，皮层脑电图可直接获取高伽马能量（HGE，70–170 Hz），该能量与发音运动学高度相关 [31], [8]。除频谱保真度出色外，皮层脑电图还具有显著的纵向稳定性 [32], [33], [9], [34]；与穿透式微电极不同，硬膜下电极阵列可减轻胶质增生，从而支持持久的慢性记录 [35]。因此，皮层脑电图已成为功能恢复性神经假体系统中连续语音重建的金标准 [9], [29], [36], [11]。

2) 神经肌肉执行：表面肌电图（sEMG）在神经肌肉接头处捕捉信号 [37]–[38]。神经解码器难以处理抽象的皮层意图，而表面肌电则能捕捉实际的肌肉激活。一项关键优势是机电延迟——电激发先于身体发音动作的 60 毫秒窗口 [29]–[38]。这种超前信号有助于实现低延迟预测解码，这对实时应用至关重要 [21]。向表皮电子学和纺织传感器的转变 [39]解决了长期以来存在的运动伪影和皮肤阻抗不稳定性问题。

B. 发音运动学

发音运动学测量声道的机械执行过程，提供一种能够抵御电磁干扰的高保真语音输出代理指标。

1) 磁性前沿：电磁发音运动分析（EMA）一直是音系学基准 [40], [41]，可提供亚毫米级三维追踪 [17], [42]。然而，有线口腔内线圈限制了 EMA 在实验室中的应用。当代技术正转向永久磁性发音运动描记（*PMA*），以推动无缆传感。通过利用

表 I：无声语音感知模态的分类法与比较分析

类别	模态	感知原理	优势与能力	技术瓶颈
神经生理（近端）				
神经生理	EEG	头皮电位	非侵入式；皮层意图。	信噪比低；颅骨衰减。
	ECoG	硬膜下振荡	高保真 HGE 接入；稳定。	手术侵入性；可及性。
	sEMG	神经肌肉	超低时延；超前。	电极漂移；运动伪影。
发音运动学（远端）				
运动学	PMA / 电磁发音运动分析	磁追踪	亚毫米精度；金标准。	有线限制；干扰。
	耳戴式设备	接近感测 / 声呐	社会接受度高；普及性强。	信息稀疏性；漂移。
	口腔内	电光	三维口腔轮廓测量；元音区分。	侵扰性高；人体工学。
成像与光学传感				
成像	实时磁共振成像	磁共振	完整的声道几何形态；鼻音性。	不可便携；成本高。
	UTI	颏下超声	可呈现舌部动态；安全。	声影；错位。
	视觉语音识别 / 深度	光学 / 彩色-深度	硬件普及；操作直观。	对光线敏感；隐私。
声学和射频（*RF*）				
声学/射频	主动声呐	调频连续波/脉冲	抗噪性；微米级跟踪。	多径干扰。
	神经声学测量 / 惯性测量单元	组织传导。	抗噪性强；可捕获亚声语音。	低通滤波效应。
	毫米波	微多普勒	**非接触**；可穿透口罩。	对非语音运动敏感。

注：*SNR*：信噪比；*HGE*：高伽马能量；*FMCW*：调频连续波。粗体条目表示当前最先进水平（*SOTA*）前沿。

可穿戴耳机中的高灵敏度隧道磁阻（TMR）传感器 [43], , PMA 可将发音器官运动与环境噪声解耦 [43], [44]。此外，光泵磁力计（OPM）能够对表面电极无法触及的深层肌肉组织进行非接触式感知 [45]。

2）解剖协同：一个重要趋势是开发利用解剖耦合的隐形接口，例如作为发音器官与耳道之间桥梁的颞下颌关节（TMJ） [46]。利用耳道变形（*ECDM*）的系统将无处不在的耳戴式设备重新用于无声语音解码 [47], [48]。尽管面临信息稀疏，其外形尺寸使其成为消费级无声语音接口的候选方案。当前研究通过振动声学融合应对这一问题，将接近传感器与骨传导惯性测量单元（IMUs）配对，以滤除伪影 [49]。

3）口腔内传感：口腔内传感对于实现最大语音覆盖仍不可或缺，其技术正从电容式腭电图（EPG） [50], [51] 向混合式电光口腔成像（*EOS*） [52], [53]发展，以解决元音盲区。传统 EPG 仅限于接触动态；通过将光学距离传感器集成到假腭中，EOS 可提供三维口腔轮廓。对于需要高分辨率区分辅音和元音的音素转文本系统而言，这些架构至关重要；当耳戴式设备保真度不足时，它们也可作为关键的临床工具。

C. 成像与光学传感

成像模态可直接呈现声道几何形态，在免受电磁干扰影响的同时，也面临固有的可观测性–普及性悖论。

1) 超声舌成像（*UTI*）：UTI 使用 B 型超声（2–10 MHz）来呈现舌部内部动态 [54], [55]。颏下换能器发射声波，声波在舌–气界面处反射，形成代表正中矢状面舌形状 [56], [57], [58]的特征性亮色轮廓。从方法学角度看，该领域已从线性基于主成分分析的表征发展至生成式神经流形。空间变换网络（STN）日益广泛地用于通过实时仿射自适应纠正探头错位 [59], , 而去噪卷积自编码器（DCAE）则可减轻复杂齿龈音发音过程中下颌遮挡所引起的声影 [55], [60]。

2）实时磁共振成像（*rtMRI*）：实时磁共振成像可全面呈现声道动态，包括软腭等关键结构，从而能够超越声学信息，对鼻音性和协同发音进行有据可依的建模 [61], [62], [63]。尽管其时间分辨率有限（通常为 5–50 Hz），rtMRI 日益成为知识蒸馏框架中的教师模型。高分辨率发音序列可用于监督深度网络，而网络所学习到的解剖结构表征则可迁移至轻量级学生传感器（例如耳戴式设备），从而在可观测性降低的情况下实现无声语音接口的实际部署。

3）视频唇读：视觉语音识别（VSR）通过利用普通摄像头捕捉可见的发音线索 [64], [65], 模拟人类唇读。特征表示涵盖基于形状的几何描述符、基于外观的感兴趣区域（ROI）像素，以及深度时空模型（例如 CNN–LSTM）。尽管 VSR 直观上颇具吸引力，但它容易受到光照变化和隐私问题的影响。这促使人们转向深度传感和可穿戴红外（*IR*）系统 [66],, 这些系统能够推断面部几何结构或颏下运动，而无需暴露完整画面的视觉内容。

D. 声学与无线电传感

声学与射频（RF）传感将声道视为一个动态的反射环境。

1) 主动声呐与射频微多普勒：主动声学传感将硬件改造为声呐系统，发射

近超声脉冲（16–22 kHz）以追踪微米级皮肤形变 [67]。时移调频连续波载波的进展解决了多用户干扰问题，使共享空间中的无设备无声语音接口成为可能 [68]。同样，毫米波（mmWave）雷达（77–120 GHz）能够透过衣物或口罩检测发音器官塑形的微多普勒特征。通过步进频率连续波（SFCW）系统采集声道传输谱 [69]，可应对时空消歧方面的挑战。

2）内部传导信号：非可听喃语（NAM）麦克风在语音产生的位置采集语音，但会受到组织低通滤波效应的影响。突破性进展涉及跨模态蒸馏：利用基于声学或实时磁共振成像数据训练的教师模型，指导处理 NAM 或惯性测量单元信号的学生模型。这样可以重建缺失的高频语音音素潜在空间，即使传感器布置发生变化，也能实现不依赖于用户的稳健性 [43]。

E. 总结与未来展望

无声语音接口研究的发展体现了研究重点从孤立探索单一模态转向协同多模态融合。每种感知范式都力求在可观测性与普及性之间的权衡中找到合适的运行点。未来的进展取决于跨模态知识蒸馏和边缘高效学习，从而利用黄金标准实验室信号，为注重隐私、稳健且为社会所接受的通信接口提供参考。

III. 无声语音接口的机器学习方法

现代无声语音接口的计算核心，是将非平稳生物信号 $\mathbf{X} \in \mathbb{R}^{C\times T}$ 高保真地转换为离散语言标记或连续声学流形。随着该领域日趋成熟，主要研究挑战已从启发式模式识别转向端到端（E2E）神经潜在对齐 [99], [100]。这一发展方向着重解决零散生理动作与序列语义单元之间的时空摩擦，反映出该领域更广泛的转变：从手工设计的启发式方法转向针对跨模态不变性优化的表征学习。

A. 无声语音接口的特征表示演进

解码后端的效能，根本上取决于能否提取出稳健的发音嵌入，使其能够应对环境非平稳性和受试者间解剖差异（见表 II）。

1）启发式基础与经典统计流形：在无声语音接口研究的奠基时期，特征提取着重于通过基于预定义规则手工设计的描述符，捕捉发音的生物力学和统计特性。对于表面肌电图（sEMG），时域（TD）特征——包括均方根（RMS）、平均绝对值（MAV） [101], 和波形长度（WL）——至今仍是标准方法，因为它们能够直接量化运动单位募集强度，且无需承受谱变换带来的计算延迟 [102]。对于高维成像模态，例如超声和实时磁共振成像（rtMRI） [103],，主成分分析（PCA）作为一种先驱技术，可将相关变量转换为互不相关的主成分 [104]。这使研究人员能够将复杂的发音轮廓 [62], [105]投影到低维子空间中，生成可解释的表征，例如特征舌 [84]或特征唇 [106]。与此同时，研究人员采用伽柏滤波器组和小波，将发音图像表示为平面正弦波，从而提取局部特征 [107]。随着这些方法日趋成熟，离散余弦变换（DCT） [87] 因具有更强的能量集中能力和更高的计算效率，成为优于PCA的替代方案。

除物理描述符之外，经典统计流形为信号净化提供了主要框架。独立成分分析（ICA）对于盲源分离至关重要，可从 EEG 和 ECoG 数据流中分离出语音相关神经意图 [108],，并排除干扰性的生物伪影。此外，线性判别分析（LDA）常用作提取后的优化方法；通过最大化类间方差与类内方差之比，LDA 可优化语音可区分性，确保即使是低资源模型也能辨别细微的发音对比[77]。

2）自主表征与深度潜在学习：深度学习（DL）的引入开创了“无特征”范式，神经网络通过直接从原始数据中学习层次化表征，将特征提取过程内化。深度自编码器（DAE） [109]因能够学习未标注数据的紧凑表征而备受瞩目，即使在大幅降维的情况下，也能保持较高的视觉保真度 [110], [55]。卷积神经网络（CNN）构建了逐步提取更高阶特征的层次结构，在模式识别方面优于传统的人工设计方法。

具体而言，二维卷积神经网络通过识别局部发音边缘和纹理线索，有效取代了传统的 DCT/PCA [54]。为捕捉动态发音轨迹，三维卷积神经网络（3DCNN）通过三维卷积核显式引入时间维度来捕捉运动 [86], [111]。随后发展出序列卷积自编码器（Sequential CAE），利用长短期记忆模型将整个图像序列处理为定长向量 [60]。为应对固有的成像噪声和散斑伪影，研究人员开发了去噪卷积自编码器（DCAE），在过滤干扰的同时保留空间信息 [112]。

3）自监督与多模态学习：近期进展的重点已转向克服标注数据长期匮乏的问题。基于掩码建模的自监督学习（SSL）范式，使模型能够借助视觉 Transformer（ViT）架构，直接从未标注数据中推断高维抽象特征 [85]。

当前最先进的解码器利用 Transformer 和 Conformer 架构中的自注意力机制，来处理协同发音中固有的复杂时间“涂抹”现象 [21]。与传统的帧级特征不同，这些模型会根据嵌入表示的全局语言重要性对其加权，从而捕捉连续语音产生的流体动力学。

表 II：神经肌肉活动连续谱中机器学习方法与性能基准的综合综述

模态	架构	数据集／语料库	词汇量	指标	结果	人群	Ref.
神经生理学							
犹他阵列	基于 Transformer	BrainGate2 (T15)	125k	词准确率	99.2%	肌萎缩侧索硬化症 / 构音障碍	[35]
犹他阵列	深度神经网络	句子解码	125k	Acc.	97.5%	四肢瘫痪 / 肌萎缩侧索硬化症	[70]
犹他阵列	RNN	BrainGate2 (T12)	125k	Acc.	96.5%	肌萎缩侧索硬化症 / 构音障碍	[35]
ECoG	序列到序列 + 大语言模型	运动学解码	连续	WER	3.0%	健康 / 临床	[8]
ECoG	卷积神经网络 (InceptionTime)	临床 (Corticom)	14 条命令	平均准确率	97.1%	肌萎缩侧索硬化症参与者	[33]
脑电图 / 肌电图	Conformer +HTNet	日语语料库	64 个单词	词准确率	95.3%	临床患者	[71]
EEG	堆叠分类器	KaraOne / FEIS	4–16	Acc.	75–86.2%	沟通障碍	[72]
EEG	大型脑语言模型	12词集	12个词	Acc.	77.3%	健康受试者	[12]
EEG	LSTM	SSDC	6个无声指令	准确率, F1分数	22.79%	10名参与者	[73]
sEMG	多任务学习Transformer	高密度表面肌电语料库	11 条命令	Acc.	98.7%	健全/假肢	[74]
sEMG	DNN-HMM混合模型	WSJ0 5k语料库	5,000	WER	6.45%	健康受试者	[75]
sEMG	HMM-SGMM（fMLLR）	喉切除术	2.5k	WER	8.9–13.6%	喉切除术患者	[76]
sEMG	深度神经网络+线性判别分析	自定义数据集	5000条话语	客观	-	健康受试者	[77]
sEMG	ResNet+Transformer	自定义数据集	-	WER	0.2696	-	[78]
sEMG	深度神经网络 + 线性判别分析	自定义数据集	5000条话语	客观	-	健康受试者	[77]
sEMG	支持向量机、K近邻和动态时间规整	自定义数据集	10个数字	WER	0.343	3位说话者	[79]
sEMG	DT	EMG-UKA试验语料库	50个单词	词准确率	95.55%	4位说话者	[80]
发音动作（超声舌成像）							
UTI（颏下）	Conformer + bi-LSTM	TaL80 / TaL语料库	1000条话语	MCD	3.03 分贝	多说话人	[81]
UTI	基于ViT	儿童SSD数据集	句子	Acc.	88.9%	儿童（言语声音障碍）	[57]
UTI	扩散/ 生成对抗网络	TaL80 / 普通话	话语	MCD	5.21 分贝	健康受试者	[82]
UTI	DNN-HMM与深度卷积自编码器	2010挑战赛	-	WER	6.17%	健康受试者	[55]
UTI	DNN + HMM	1023词集	1,023	词准确率	94.1%	健康受试者	[83]
UTI	主成分分析 + 多层感知机	IEEE/哈佛	-	NMSE（LSF）	11–16%	健康受试者	[84]
UTI	基于视觉Transformer的自监督	UXTD	4609 句子	Acc.	13.33%	健康受试者	[85]
UTI	三维卷积神经网络 自监督	WSJ0 / 天津大学超声舌成像	735k / 147k	MSE	21.7-32.6	健康受试者	[86]
UTI	隐马尔可夫语音模型（基于离散余弦变换）	视觉TIMIT / WSJ0	5k / 20k	词准确率	69–79%	健康受试者	[87]
UTI	序列卷积自编码器	2010挑战赛	-	WER	5.96%	健康受试者	[60]
UTI	KNN	自定义数据集	30名德国人	准确率	52%	单说话人	[88]
UTI	基于 Transformer	WSJ0 2k 语料库	单词 2000	字符错误率，词错误率 字符错误率 = 10.1%, 词错误率 = 20.5%		单说话人	[89]
UTI	自编码器	自定义数据集	句子	NMSE	-	单说话人	[90]
运动学及其他							
耳戴式设备（MxG）	深度学习 / 岭分类器 Sonera 定制		12 条命令	Acc.	86.5%	可穿戴设备用户	[45]
EMA	双向长短期记忆隐马尔可夫模型 / 深度典型相关分析	132 个短语	39 个音素	PER	33.9%	喉切除术患者	[91]
EMA	高斯混合模型-隐马尔可夫模型	自定义数据集	37 音素	PER	47.5%	5 名英语母语者	[92]
视频（唇部）	LipNet（时空卷积神经网络 + 双向门控循环单元）	GRID语料库	51 个单词	句子准确率	95.2%	健康受试者	[93]
视频（唇部）	VALLR（视频转换器 +大语言模型）	LRS3数据集	大型	WER	18.7%	健康（数据效率）	[94]
声学与射频							
SAAS (PMUT)	ResNet (残差网络)	每日通信	10句子	Acc.	99.8%	潜在患者	[95]
毫米波	mSilent（卷积神经网络 + Transformer 序列到序列）	mSilent语料库	1k 句子	WER	9.5%	健康（非接触式）	[96]
空气传播超声	卷积神经网络 + 连接时序分类	自定义数据集	31 条命令	WER	4.5%	健康受试者	[15]
耳内超声	EarS SR（分层卷积神经网络）	自定义数据集	50 个单词	Acc.	93.0%	可穿戴式（耳内）	[97]
耳内声呐	EarCommand（CRNN +连接时序分类）	EarCommand	57 条命令	WER	10.0%	健康受试者	[14]
吸入气流	训练好的识别器	SilentVoice	80 句子	WER	5.0–6.7%	健康受试者	[98]

此外，FusionNet 等多模态方法通过将原始成像数据与利用局部二值模式（LBP）提取的纹理特征相结合，提高了分类准确率 [113]。

4) 将物理建模和生物先验作为归纳偏置：尽管数据驱动方法占据主导地位，现代无声语音接口研究已重新引入物理建模，将其作为一种严格形式的归纳偏置 [114], [115]。这种混合方法在临床场景中至关重要，因为数据稀缺限制了黑箱模型的泛化能力 [116]。一种典型范式是生物力学约束 [117], [118], ，即将主动轮廓模型（蛇形模型）嵌入神经网络流水线，以强制满足解剖学合理性 [58]。通过将舌部跟踪限制在生理有效边界内，这些模型能够防止在解码过程中“臆造”出不可能的发音

器官状态 [39]。

此外，计算机视觉与肌肉骨骼建模的融合催生了无标记运动学提取技术。DeepLabCut 等框架能够从非侵入式视频或超声数据流中提取“虚拟标记” [119], [120], , 在无需承受侵入式电磁正中矢状面发音运动分析（EMA）传感器操作不适的情况下，模拟其精确度。与原始像素或肌电数据相比，这些运动学特征对个体生理外观具有更强的内在不变性，从而促进更优的跨说话人迁移学习，并应对长期存在的说话人无关校准难题 [43]。

B. 无声语音识别的发展

无声语音识别（SSR）的算法格局已经完成了从基于启发式的统计建模向高容量端到端神经网络架构的根本性转变 [78]。这一演变体现了该领域日益增强的能力：解析随机生物信号与离散语言单位之间的非线性映射。现代解码框架的特点在于能够建模复杂的时间依赖关系、提取不变的空间特征，并缓解数据稀缺与跨个体解剖结构差异所带来的普遍挑战。

1）统计基础与启发式分类器：早期的无声语音识别系统主要建立在经典统计框架之上，旨在进行孤立词识别或受限识别。隐马尔可夫模型（HMM）历来是系统架构的核心 [121], [54], [92], , 通常与高斯混合模型（GMM-HMM）结合，用于表征发音状态的频谱或肌电特性。尽管这些模型为基准数据集奠定了基础，但由于建模无声语音非平稳动态特性的能力有限，往往呈现较高的错误率。动态时间规整（DTW）等互补技术被用于模板匹配，尤其是补偿运动学传感器数据中的时间失真 [79],，但其计算复杂度限制了它们在连续语音任务中的应用 [88]。在低资源场景中，支持向量机（SVM）和随机森林等判别式分类器仍适用于从小规模脑电图或表面肌电数据集中识别音位单位，而深度学习在此类场景中仍容易过拟合。例如，有研究表明，将TD6时域特征与决策树分类器结合，在通道数从七个减少到五个的同时，仍能对表面肌电信号实现闭集识别，且性能未显著下降 [80]。

2）神经表征与时空建模：深度神经网络（DNN）的集成标志着一种范式转变，即开始捕捉神经肌肉数据和成像数据中固有的潜在非线性。一个早期里程碑是 DNN-HMM 混合模型，它以深度架构取代高斯混合模型，用于发射概率估计，从而显著降低了基于超声的系统中的词错误率（WER） [54], [89]。为进一步利用高密度表面肌电网格或视觉音素的空间结构，研究人员采用卷积神经网络作为分层特征提取器。尤其是，3D 卷积神经网络已成为视频和超声序列中不可或缺的工具，它们在空间和时间维度上进行卷积，以捕捉发音轨迹的流体动力学特征 [57]。为建模语音的序列特性以及协同发音的复杂影响，长短期记忆（LSTM）和双向长短期记忆网络（Bi-LSTM）等循环架构已成为标准方法，它们利用过去和未来的上下文来消解语音歧义 [90]。

3）端到端对齐与注意力机制：最先进的静默语音识别解码器已广泛转向端到端（E2E）学习框架，无需中间的语音对齐步骤，直接将原始生物信号映射为文本。在此背景下，连接时序分类（CTC）已成为一种关键损失函数，为将不受约束的发音信号与长度各异的目标文本序列对齐提供了数学解决方案。它通常与序列到序列（Seq2Seq）编码器-解码器架构配合使用，该架构利用注意力机制，在解码过程中智能地为相关发音帧分配权重 [73], [96]。目前，融合了卷积神经网络的局部特征提取能力与Transformer的全局上下文建模能力的Conformer架构，代表了当前的性能前沿；它能够捕捉细粒度的运动学细节和高层语义依赖关系，在各种模态上均取得了更优的结果 [20]。

4）泛化与语义优化策略：为克服“数据瓶颈”以及用户之间的解剖差异，研究人员采用了复杂的鲁棒性策略。他们运用SpecAugment和时间扭曲等数据增强技术，人工扩充有限的数据集 [45]。迁移学习和领域适配（例如领域对抗训练）能够对齐不同受试者的特征，促进“说话人无关”系统的开发，减少重新校准的需求。此外，通过Procrustes匹配或特征空间最大似然线性回归（fMLLR）进行生理归一化，有助于缩小个体之间的解剖差异。另一项重要进展是将大语言模型（LLM）集成到后处理环节：根据语义概率对静默语音识别解码器生成的候选输出进行重评分，从而有效纠正语音幻觉，并显著降低最终词错误率 [20]。

C. 直接无声语音合成：从发音到声学

发音到声学（ATA）范式，又称直接合成，是静默语音接口研究中的一个成熟分支，旨在将生理生物信号直接映射为声学波形。与依赖中间文本解码的识别型系统不同，ATA 框架采用连续映射函数。这种方法显著降低了处理延迟，并保留了副语言细微特征，例如个性化的嗓音身份和情感韵律，而这些特征在文本约束流程中往往会被舍弃 [90]。如表 III 所示，该领域已从僵化的统计映射发展到能够实现高保真频谱重建的生成式架构。

表 III：直接无声语音合成的基准数据集：发音到声学（ATA）范式

传感器类型	模型架构	数据集／语料库	词汇量	指标	结果值	人群	Ref.
神经生理							
神经营养电极	卡尔曼滤波神经解码器	元音发音	3 个元音	命中率	70.0%	闭锁状态	[36]
ECoG	双向长短期记忆网络（序列到序列）	连续句子	Open	MCD	4.66–5.07 分贝	癫痫	[122]
运动学							
sEMG	DiffMV-ETS（扩散）	EMG-VCTK	句子	字符错误率 / 归一化平均意见分	45.8% / 3.0	健康	[123]
sEMG	深度神经网络 / 高斯混合模型转换	Arctic / TIMIT	108 个单词	词错误率 / 梅尔倒谱失真	7.3% / 4.5 分贝	健康	[77]
sEMG	纹身式贴片 + 线性判别分析	指令	11 条命令	Acc.	89.0–92.3%	Lar.	[124]
sEMG	Transformer +WaveNet	句子（书籍）	Open	WER	42.2%	健康	[125]
PMA	循环神经网络（门控循环单元）/ 多模态融合	Arctic / TIDigits	48 个项目	准确率 / 梅尔倒谱失真	91.5% / 4.9 分贝	健康	[17], [126]
成像							
UTI	Conformer +HiFi-GAN	Ultrasuite-Tal80	728 话语	MCD	3.03–4.13 分贝	健康	[81]
UTI	卷积神经网络 +WaveGlow	TaL / PPSD	Open	MCD	5.37 分贝	健康	[127]
UTI	DDPM（扩散）+平行波生成器	TaL80	话语	梅尔倒谱失真 / 词错误率	3.53 分贝 / 73.3%	健康	[128]
超声舌成像 + 视频	全协方差	法语句子	Open	准确率 / 梅尔倒谱失真	60% / 5.2 分贝	健康	[129]
超声舌成像 + 光学	隐马尔可夫模型/高斯混合模型多层感知机映射 (Eigentongue)	IEEE/哈佛	720 句子	NMSE	11.0%–16.0%	健康	[84]
声学/射频							
超声	U形网络 +Transformer + PWG	LJSpeech / TIMIT	N/A	感知语音质量评估 / MOS	3.10 / 4.48	健康	[130]
神经声学测量 / 电子喉	高斯混合模型语音转换 (VC)	ATR 音素	140 句子	梅尔倒谱失真 / MOS	4.65 分贝 / 3.8	健康/喉切除者	[131]

1）统计基础与早期神经映射：对于神经生理信号，早期皮层内系统采用线性状态空间解码器奠定了基线，其中卡尔曼滤波器将低维神经轨迹映射为声学参数，以供闭锁综合征患者使用 [36]。随着该范式扩展至运动学和成像模态，研究人员改造了统计语音转换框架，主要采用高斯混合模型（GMM）来估计发音到频谱映射的联合概率密度 hirahara2010silent。为确保声学连续性，随后采用基于隐马尔可夫模型的生成模型施加明确的时间约束，预测平滑的参数轨迹（如梅尔倒谱轨迹），而非彼此独立的帧 [129]。尽管这些统计方法发挥了奠基性作用，但往往难以处理语音产生中复杂的非线性动态。这一局限在高维超声舌成像数据中尤为明显，并促成了具有里程碑意义的 Eigentongue 分解方法；该方法应用主成分分析，将原始超声图像投影到紧凑的发音子空间 [84]。向紧凑表示的这一演进推动了深度神经网络和“瓶颈”架构的采用 [77]。这些多层感知机通过促使网络内化发音输入中富含信息的潜在表示，显著提高了映射准确率，并为现代生成式合成奠定了基础。

2）序列建模与生成对抗网络：为解决协同发音带来的挑战（即声学输出取决于周围的发音状态），序列感知模型逐渐成为标准架构。Bi-LSTM 网络和门控循环单元（GRU）已成功用于解码神经生理学和运动学信号中的复杂时序依赖关系 [122], [17]。当前的高性能骨干网络进一步采用 Transformer 和 Conformer 层，利用全局自注意力，实现更出色的图像与声学转导 [81],[130]。为缓解传统回归固有的“过度平滑”效应，合成流程中采用了生成对抗网络（GAN）。具体而言，集成 HiFi-GAN 和 Parallel WaveGAN（PWG）等神经声码器，将中间频谱特征转换为高保真可听波形 [81], [128], [130]。

3) 扩散模型与神经声码：当前 ATA 合成的技术前沿已转向去噪扩散概率模型（DDPM） [128]。*Diff MV-ETS* 等架构通过逐步逆转高斯加噪过程，将信号从纯噪声迭代细化为高保真梅尔频谱图，并以发音输入为条件（直接或间接） [123]。由于扩散模型生成的声学特征相比以往的非概率映射模型较少过度平滑且更多样，因此越来越受青睐 [128], [123]。至关重要的是，扩散模型会先预测一种中间频谱表示，再由声码器将其转换为可听波形 [128]。早期的流程通常依赖参数化／统计声码（例如预测 MFCC/F_0 并使用 MLSA），这需要基频等明确的激励相关特征（F_0）[77], ；现代框架则越来越多地采用 WaveNet [125], 、WaveGlow [127],或 HiFi-GAN [81] 等神经声码器。这些神经解码器可补充细粒度细节，从而实现较高的可懂度和自然度。

4) 训练算法创新：一项关键挑战
在针对“静音”语音训练ATA模型时，固有的

表 IV：语言先验与大语言模型前沿：分类基准。

传感器类型	模型架构—ture	数据集／语料库	词汇量	指标	结果值	人群	Ref.
神经生理							
MEA	循环神经网络 + n元语法 LM + LLM 重评分	提示式句子	12.5万词	WER	2.5%	肌萎缩侧索硬化症患者	[132]
EEG	Transformer + 残差块 + CTC	《爱丽丝》有声读物	601 个单词	前十准确率	26.82%	健康	[31]
运动学							
表面肌电 +音频	MONA LISA (GPT-3.5 微调-调)	Gaddy 基准	Open	WER	12.2%	受损	[20]
sEMG	Transformer + GPT-2 后处理	数字 发声（Gaddy）	-	WER	30%	-	[133]
sEMG	隐马尔可夫模型-深度神经网络 + 组合语音特征	-	-	WER	34.7%（静音）	健康	[8]
面部表面肌电	Transformer + CTC + Aux. 任务	NBA 肌电图到 Text	667个字符	CER	38.0%	健康	[74]
sEMG	肌电信号适配器 + Llama3-3B	Gaddy 和 Klein (2021)	67 个单词	WER	0.49	健康	[134]
肌电图 + 压电陶瓷	孪生网络 (孪生神经网络) + PMLDF	字母	8 个字母	Acc.	95.63%	健康	[135]
成像							
视频（唇部）	LipType（LLaMA-3.2-3B）	LRS3测试集	大型	WER	9.19%	健康	[94]
视频（唇部）	Auto-AVSR	LRS3测试集	大型	WER	20.3%	健康	[94]
超声	基于ViT 自我-有监督的	Eshky et al. (2018)	-	Acc.	88.94%	儿童	[57]
超声	基于卷积神经网络的自-监督	-	-	Acc.	79.4%	健康	[57]
声学/射频							
超声 + 视频	基于隐马尔可夫模型 + HNM	CMU ARCTIC	3000个词	Acc.	83.3%	健康	[136]

缺少真实音频目标。为解决这一问题，研究人员开创了伪目标生成方法。系统最初使用发声数据进行训练，其中发音动作信号和声学信号共存；随后，无声发音动作数据会与通过 DTW 等对齐算法生成的“幻觉”音频配对，或与通过文本转语音（TTS）引擎合成的音频配对。这种合成监督使 ATA 模型即使在缺少原始发声录音的情况下也能提升其映射能力，推动其从实验室原型走向实用的临床应用 [128],[123]。

D. 语言先验与大语言模型前沿

从历史上看，无声语音接口的计算瓶颈一直在于生物信号固有的歧义性和非平稳性。近期，大语言模型被用作无声语音接口系统的语义引擎，推动了范式转变，如表 III 所总结。现代框架不再仅依赖声学或发音动作建模，而是利用大语言模型丰富的语言先验进行潜在语义对齐，将零散的生理动作转化为连贯的交流。

1）功能分类：大语言模型不断演变的作用：大语言模型在无声语音接口流程中的功能集成，已从外围后处理发展到核心信号解码。最初，大语言模型充当高层次的“语义修正器”。例如， [133], [20] 提出了一种双阶段框架：先由基于Transformer的自动语音识别将信号转换为噪声文本，再由大语言模型解决句法不一致和语音歧义。更为复杂的方法以MONA LISA系统 [20],为代表，它利用大语言模型集成评分调整（LISA），对束搜索得到的 N个最佳候选结果重新评分，优先考虑语义合理性，而非原始信号似然度。除了纠错之外，近期研究还探索了直接信号到语义翻译，将生理特征视为“神经令牌”。 [134]展示了一种肌电信号适配器，可将表面肌电特征直接映射到Llama-3主干模型的嵌入空间，使模型能够将肌肉活动解读为一种高维方言。这一转变使零样本/少样本个性化成为可能，从而应对个体化数据长期匮乏的问题。**LipLearner** 系统 [137] 利用语言迁移学习，仅凭六分钟训练数据便实现了个性化指令识别。此外，自然语言提示现已用于对大语言模型进行特定模态的条件设定；

例如，采用 $UnvoicedEMG{:}<emg_embed>Prompt{:}Converttotext$ 等特定前缀提示，引导模型关注发音动作特异性潜在表征 [134]。

2）架构趋同：对齐与融合机制：主要的技术挑战仍然是模态对齐问题——将连续、高频的生物信号投射到离散的语义潜在空间中。当前最先进水平的系统采用基于一维卷积神经网络或 ResNet 的适配器，对表面肌电或脑电图等高频信号进行下采样，并将其投射到大语言模型的嵌入维度。为弥合“无声鸿沟”，交叉对比损失（$crossCon$）和监督式时间对比损失（$supTcon$）会在共享潜在空间中对齐无声信号流形与其有声对应信号。架构趋势显示，模型正逐渐趋向混合架构，通常将 Conformer 视觉/信号编码器与 Llama-3.2（1B 或 3B）解码器结合，以平衡性能与移动边缘部署需求。为了在适应小众生理数据的同时保留大语言模型的海量知识，参数高效微调（PEFT），尤其是低秩适配（LoRA），已成为标准方法，能够在不导致底层语言基础知识“灾难性遗忘”的情况下，实现较高的无声语音接口准确率 [94]。

3）性能基准与可用性阈值：大语言模型的集成使词错误率实现了历史性下降，尤其是在开放词汇场景中。MONALISA系统取得了一项重要里程碑：在Gaddy（2020）基准上，将词错误率从28.8%降至12.2%。这是非侵入式表面肌电无声语音接口首次突破实际应用所需的15%“可用性阈值”。然而，准确率的提升也带来了显著的延迟成本。尽管Llama-3.2-1B等模型针对边缘设备进行了优化，但使用大型大语言模型（如GPT-4）进行重评分会造成显著的推理延迟，这凸显了语义深度与实时交互之间的关键权衡。

4）语义鸿沟与未来前沿：尽管取得了这些进展，“语言学前沿”仍面临两项关键挑战：数据鸿沟与幻觉瓶颈。前者指的是万亿级词元文本语料库与小时级生理数据集之间的根本性不匹配。当前方法严重依赖平行的“有声”数据作为桥梁，但失声患者往往无法提供这类数据。后者指的是过度纠正的风险：如果输入信号受损过于严重，大语言模型可能会“幻觉”出一个语法完美、却完全偏离用户实际意图的句子 [21]。近年来，研究趋势正转向大型脑语言模型（LBLMs） [12]。这标志着研究思路从“改造大语言模型”转向“构建原生生理大语言模型”，即直接使用大规模自监督生物信号数据集对模型进行预训练。这种方法旨在解决对齐问题：不再将人类神经活动视为有待翻译的含噪输入，而是将其视为一种原生的语言模态。

展望未来，发展趋势正转向大型脑语言模型（LBLMs）。这标志着研究方向从“改造大语言模型”转向构建“原生生理大语言模型”这类模型直接在海量自监督生物信号数据集上进行预训练。该方法旨在解决对齐问题：它不再将人类神经活动视为充满噪声、需要翻译的输入，而是将其视为一种原生的主要语言模态。通过将重建负担从噪声较大的物理传感器转移到稳健的语言先验上，该领域终于接近了 15% 的词错误率门槛，使无声语音接口从实验室里的新奇事物发展为切实可行的辅助技术。

IV. 基准测试：评估指标与开放数据集

无声语音接口迈向可复现开发，有赖于标准化的评估架构。本节建立了一个评估框架，用于在识别与合成模态上评估无声语音接口，从而弥合原始信号采集与交际效用之间的差距。

A. 定量性能指标

评估无声语音接口需要多维视角，综合考量语言精确度、声学自然度与运行效率。由于这些接口将非声学生物信号（如表面肌电、超声或脑电图）映射到高层语义单元，简单的准确率指标往往无法捕捉人在回路体验中的细微差别 [80]。因此，研究人员综合运用信息论与心理声学模型，以衡量临床队列的恢复效能 [20], [138]。

1) 语言准确度（识别指标）：词错误率（WER）仍是基于识别的解码器的首要基准，由莱文斯坦距离得出：

$$\text{WER} = \frac{S + D + I}{N} \times 100 \tag{1}$$

其中 S, D, I 分别表示替换、删除和插入， N 为参考计数 [96]。词错误率低于15%阈值通常被视为无声语音接口实现与传统自动语音识别功能等效的必要条件 [20]。鉴于无声发音的语音密度，研究日益采用音素错误率（PER）或字符错误率（CER）来消解细粒度歧义，尤其针对词表外（OOV）术语以及普通话等声调语言 [139], [140]。除错误率外，通信吞吐量由信息传输率（沃尔帕速率）量化，当前最先进的神经解码器已接近每分钟100词的自然会话速度 [32]。

2) 声学与感知保真度（合成指标）：直接合成（发音到声学）系统结合客观信号处理与主观感知来评估重建波形。与识别不同，发音到声学旨在通过建模非线性发音到声学流形来保留韵律线索 [90]。具体而言，梅尔倒谱失真（MCD）是衡量频谱相似度的标准客观指标：

$$MCD[dB] = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (c_d - \hat{c}_d)^2} \tag{2}$$

其中 D 表示特征维度。尽管针对 UTI 或 EMA 信号的高保真系统报告的 MCD

值介于 4.5 分贝和 6.5 分贝之间，但由于统计模型的“平均效应” [129]，客观性能的提升并不总是与可懂度相一致。为更好地近似人类听觉，研究人员分别采用 PESQ 和 STOI 来衡量自然度和时间包络相关性 [68]。现代框架进一步采用扩展短时客观可懂度（ESTOI），以保留频谱回归中丢失的语音对比。此外，由于声带振动与声门上运动之间缺乏直接的解剖学联系，恢复韵律保真度仍是一项重大挑战 [57]。当前系统利用辅助模块预测 F_0 轨迹以及浊音/清音（V/UV）状态，以避免机械单调[141]。最后，平均意见分（MOS）仍是主观验证的金标准 [142]，，近期则转向 MUSHRA协议，以针对隐藏锚点实现更高的统计敏感性 [90]。

3）系统级约束：实际部署需要以推理时延和实时因子（RTF）为基准进行评估。为实现无缝交互，端到端延迟必须低于 50 毫秒，以避免言语不流利[143]。可穿戴设备研究还优先考虑功耗（毫瓦）和计算复杂度（浮点运算次数）；如今，低功耗平台已能将采集功耗维持在 22.2 至 73.3 毫瓦范围内 [144], [15]。

B. 开源数据集与资源库

从封闭集实验室实验转向大规模、开放词表解码，得益于多个关键多模态资源库的发布 [145]。这些数据集针对无声语音接口的核心挑战，例如“无声隆巴德效应”，以及发声模式与静默模式之间的构音失配 [146]。在表面肌电资源中，Gaddy数据集（19小时）通过平行的无声/有声录音，仍是数字发声研究的基石 [125]。在成像领域，无声语音挑战档案提供了超过70万帧同步的超声和唇部图像 [147]，，而LRS3-TED则是视觉语音识别领域的标准“真实场景”基准，包含超过400小时的数据 [148]。近期进展还带来了大规模普通话数据集和基于神经信号的数据集。AVE Speech数据集提供了来自100名参与者的55小时同步音频、视频和6通道肌电图数据，这对研究声调语言至关重要 [149]。在神经信号领域，MindSpeak 2025资源库（120小时高密度脑电）有助于大型脑语言模型（LBLM）的自监督预训练 [150]。与此并行，EchoSpeech在眼镜设备上采用主动声学传感，以73.3 mW的低功耗实现了高准确率的构音追踪 [15]。SSR7000等专门资源库关注“实验室外”的多样性，提供原始数据，用于评估模型应对传感器错位的鲁棒性[145]。MultiNAM数据集和ReSSInt项目支持临床修复应用，借助高保真神经声码器开展喉切除术和肌萎缩侧索硬化症康复，保留说话人身份和表现性韵律 [151], [143]。

C. 标准化与泛化难题

尽管开放源代码资源库不断扩充，无声语音接口的稳健标准化仍受到若干根本性瓶颈的阻碍。

首先，临床和健康受试者群体之间的解剖异质性——包括舌形态和神经肌肉组成的差异——会导致生物信号发生显著的分布偏移。这种受试者间变异使零样本、受试者无关迁移持续面临挑战，因此需要采用复杂的领域自适应策略来对齐差异显著的用户流形 [166], [57]。其次，系统必须应对跨模态与会话不匹配。在静默模式下，缺少听觉反馈往往会引发“无声隆巴德效应”，表现为发音动作力度减弱和构音不足。此外，可穿戴传感器对不同会话间亚毫米级的位置变化高度敏感，导致严重的信号失准，使解码模型的纵向稳定性更难保证 [83], [167]。最后，由于缺乏将高维成像与离散语义单元同步的通用协议，标注与对齐鸿沟仍然存在。将连续超声序列映射到声学或语言特征需要达到亚毫秒级精度，但发音动力学与目标音位之间的信息不对称往往会导致时间抖动和解码伪影 [168], [130]。

V. 无声语音接口的应用

无声语音接口日趋成熟，其用途已超越无声学识别。通过将语言表达与声带振动解耦，无声语音接口推动了信息外化方式的根本转变，这种转变可归纳为临床功能恢复、隐蔽交互和普适计算。

A. 临床康复与表达功能恢复

静默语音接口研究的首要目标，是为患有严重发声障碍的人恢复沟通自主能力。

1) 喉切除术后的语音康复：全喉切除术会使患者失去主要的声音来源，往往导致社交孤立。传统的无喉发声方法（如电子喉）常因声音听起来机械且社会接受度低而受到批评 [169]。无声语音接口提供了更优的替代方案，可将完好的构音控制直接映射为个性化语音合成 [8]。现代系统利用声音银行驱动神经声码器，以恢复用户独特的音色和嗓音身份，延迟低于50毫秒——这对于维持说话人的听觉反馈环路至关重要 [170], [76]。

2）神经运动增强：对于肌萎缩侧索硬化症或闭锁综合征患者而言，传统的辅助与替代沟通（AAC）设备受限于较低的吞吐率（< 10 每分钟词数）和较高的认知疲劳程度 [34]。无声语音接口通过解码默读构音意图，实现高带宽通信。高分辨率皮层内脑机接口已实现对话吞吐率超过每分钟 60 词，准确率达到 $> 99\%$ [171]。先进的基于皮层脑电图的解码器

表 V：无声语音研究代表性开源数据集汇总。

数据集	模态	受试者	词汇量	时长	主要特征	数据集来源 / 链接
I. 表面肌电图 *(sEMG)* 数据集						
EMG-UKA（完整 / 试验） [152]	表面肌电 +音频	8 / 4	2,102 /108单词	7.7 / 1.87 小时	多模式（可听、耳语、静音）；广播新闻领域。用于特征提取的公开子集基准测试。	csl.anthropomatik.kit.edu
Gaddy sEMG [19], [125]	表面肌电 +音频	1	67 个单词	18.6 小时	配对有声/无声录音实现数字发声的突破录音。	zenodo.org/records/4064409
AVE Speech [149]	音频 + 视频 + 肌电图	100	100 句子	55.3 小时	首个大规模普通话句子级肌电-视觉语料库。	huggingface.co/datasets/MML-Group/AVE-Speech
EMG2QWERTY [153]	腕部表面肌电	10+	QWERTY按键	大规模	使用可穿戴表面肌电腕带进行盲打解码任务。	huggingface.co/datasets/cyrilzakka/emg2qwerty
II. 超声舌成像（*UTI*）和视觉语音识别（*VSR*）数据集						
SSC档案库 [154], [87]	超声舌成像 +光学	1	5,000 个单词	70 万张图像	一个里程碑式的基准，提供同步的舌部超声和基于 TIMIT 与 WSJ0 的唇部视频语料库。	ftp.espci.fr/pub/sigma/
无声语音识别 7000 [145]	超声舌成像 + 光学	1	开放词汇	7,484 个话语	针对端到端SSR优化的大规模同步语料库；包括原始数据、预处理数据和特征数据。	github.com/supernaiter/SSR7000
TAL [155]	超声舌成像 + 视频 + 音频	82	开放词汇	24小时	包含一位说话人的多说话人同步语料库专业人士（TaL1）和81位普通说话人（TaL80），涵盖常规模式和静默模式。	ultrasuite.github.io/data/tal_语料库/
UltraSuite [156]	超声舌成像 + 音频	113	混合	18.7 小时	专注于儿童言语治疗的临床数据库；包含典型和言语障碍发展。	ultrasuite.github.io
AUSpeech [148]	超声舌成像 + 音频 + 文本	54	普通话句子。	22.31 小时	首个高分辨率（920×700）普通话音频超声同步数据库，涵盖所有允许的音节首音以及日常句子。	cstr.cn/31253.11.sciencedb.18722
LRS3-TED [157]	视觉（VSR）	数千种开放词汇		400+ 小时	规模最大的真实场景视觉语音数据集；核心资源用于大规模唇读预训练和大语言模型增强型模型。	robots.ox.ac.uk/~vgg/data/lrs3/
LRW-1000 [158]	视觉（VSR）	2000	500-1k 个单词	1M+ 个实例	大规模自然分布基准数据集，用于词级英语（LRW）和普通话（LRW-1000）视觉语音识别。	vipl.ict.ac.cn/view_database.php
III. 脑电图与神经信号数据集						
大声思考 [159]	神经（脑电图）	10	5 个单词	多会话	用于西班牙语内言语识别的开放获取的脑机接口数据集识别。	github.com/N-Nieto/Inner_语音数据集
Alice脑电图 [160]	神经（脑电图）	33	开放（故事）	12.4 分钟	用于语言预测的被动有声书聆听分析。	deepblue.lib.umich.edu/data/concern/data_sets/bg257f92t
SPARRKULEE [161]	听觉脑电图	85	连续	168小时	用于语音的大型听觉诱发反应资料库刺激。	rdr.kuleuven.be/数据集。xhtml?persistentId=数字对象标识符:10.48804/K3VSND
IV. 磁感测与发音成像（电磁发音运动分析/*PMA*/磁共振成像）						
MOCHA-TIMIT [162]	电磁发音运动分析 + 电腭图 + 音频 2		460 句子	920 话语	用于协同发音建模的经典发音数据库。	cstr.ed.ac.uk/research/projects/artic/mocha.html
USC-TIMIT [163]	实时磁共振成像 + 音频 / 电磁发音运动分析 + 音频	10 / 4	460 句子	3.05 小时 / 460 话语	高质量声道扫描及配套的降噪音频；与磁共振成像同步的肉点追踪数据语料库。	sail.usc.edu/span/usc-timit/
MNGU0 [164]	电磁发音运动分析 + 音频	1	开放词汇	67分钟	单说话人电磁发音运动分析数据，包含第1天/第2天子集。	mngu0.org
V. 声学、雷达及其他感知模态						
MultiNAM [151]	神经声学测量 + 视频 + 音频	2	21 话语	7.96 小时	配对的非可听喃喃语音、面部视频和耳语。	diff-nam.github.io/DiffNAM/
TAPS [165]	喉部 + 音频	60	4,000 条话语	15.3 小时	配对的声门上喉部麦克风与声学录音。	huggingface.co/datasets/yskim3271/喉部与_声学配对语音_数据集

能够提供稳定的长期性能，使瘫痪用户可以独立操作个人电脑并维持职业工作。

3）发音动作生物反馈：在言语治疗和第二语言（L2）习得中，无声语音接口可提供实时视觉—声学生物反馈。失用症等障碍会导致难以协调看不见的发音器官。通过超声舌成像（UTI）或磁追踪（EMA）将舌部运动可视化，用户可以即时观察当前发音动作与语音目标之间的偏差 [57]。这有助于重新校准运动学轨迹，提升中风或口腔癌幸存者的可懂度 [172], [173]。

B. 隐蔽交互与隐私保护型人机交互

无声语音接口可缓解声学通信的“广播性”，从而实现在对噪声敏感或安全要求高的环境中的高带宽交互；在这些场景下，传统语音用户界面（VUIs）并不实用。

1) 隐蔽数字能动性：无声语音接口通过允许用户借助内部发音——即在不发声的情况下轻微启动发音器官——与人工智能助手互动，促成了“人机认知融合” [13]。AlterEgo 等系统可捕捉残余神经肌肉信号，从而让用户在公共空间中进行双向对话而不引起

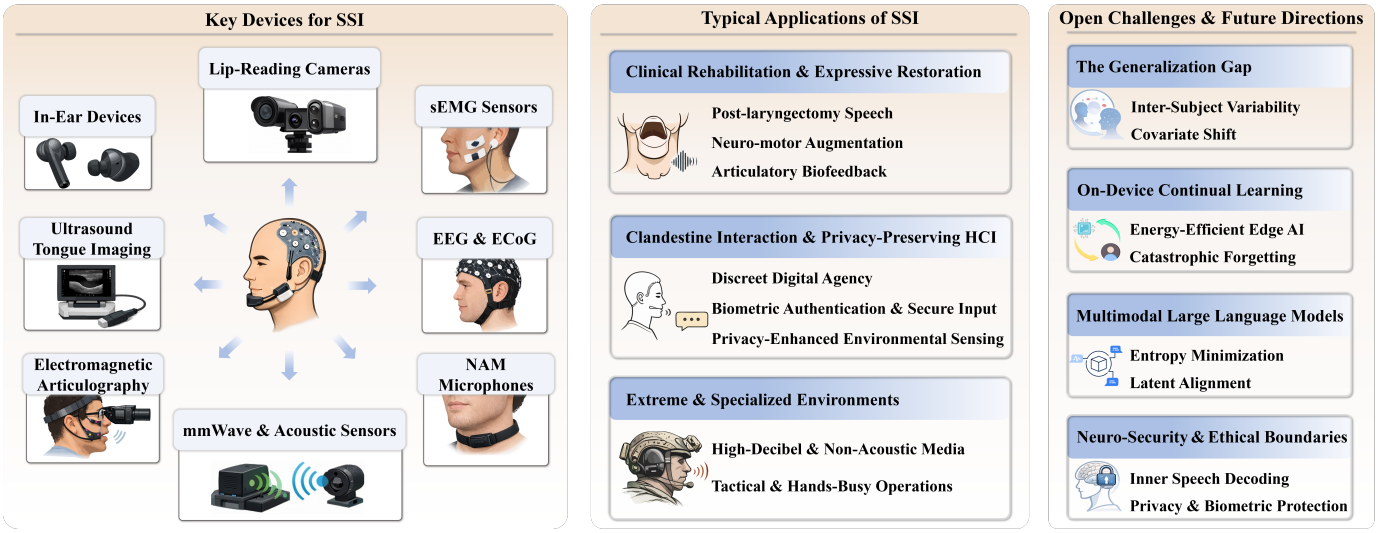


图4：无声语音接口（SSI）概览：关键传感模态、典型应用领域、开放挑战与未来方向。该分类法连接了原始生理信号截获与临床及普适计算环境中的高带宽通信应用之间的鸿沟。

旁人注意。这一范式将计算视为用户认知的自然延伸，解决了人们在公共场合以口头方式与设备交互时常有的社交尴尬问题[137], [174]。

2）生物特征认证与安全输入： 无声语音与独特的解剖特征相结合，可实现稳健的安全框架。诸如 **HEar-ID** 之类的系统会采集耳道动态运动（ECDM）作为生物特征签名 [175]，从而支持能够抵御声学窃听或重放攻击的“静默密码”。此外，**ToothSonic** 和 **Lipwatch** 等技术利用发音动作生物特征签名在输入过程中验证身份，确保只有合法用户才能访问设备 [18], [175]。

3）隐私增强环境感知： 采用毫米波雷达（*mSilent*）或主动声学传感（*EchoSpeech*）的非接触式无声语音接口，为基于摄像头的唇读提供了保护隐私的替代方案 [96]。与能够捕捉可识别面部纹理的RGB视频不同，无线传感捕捉的是粗粒度的发音动力学，可最大限度减少生物特征泄露，同时能够穿透口罩等遮挡物工作 [15]，[176]。

C. 环境鲁棒性与特殊场景

无声语音接口通过捕捉语音产生的生理成因，而非其声学后果，为传统麦克风在功能上无法胜任的环境提供了解决方案。

1) 高分贝与非声学媒介：在极端噪声环境中（信噪比 < 0），标准自动语音识别的性能会急剧下降。无声语音接口通过利用毫米波、表面肌电或非可听喃语（NAM）传感器来捕捉人体传导的振动，从而不受影响 [151]。现代皮肤附着声学传感器（SAAS）能够在 125 分贝的环境中维持信号采集，而空气传导麦克风则会失效 [95]。此外，无声语音接口还能在缺乏声学媒介的环境中实现言语交互，例如水下或舱外航天作业环境中 [177], [178]。

2）战术行动与双手繁忙的作业： 对于执行关键任务的人员（如消防员或士兵），耳戴式无声语音接口或纺织传感器无需清晰的声学通路即可实现免提控制。这对于佩戴呼吸装置或危险品防护服的用户至关重要，因为这些装备会造成严重的声学遮挡 [14]。近期的进展使多通道纺织品肌电传感器能够直接集成到耳罩中，在保持隐蔽性的同时，实现高准确率的无声控制 [179], [45]。总之，无声语音接口充当声道的“数字代理”，即使生理或情境因素限制了发声，也能捕捉机器可解读的意图耳语。

VI. 开放挑战与未来方向

尽管无声语音接口已从实验室原型发展到大众化可穿戴设备，其普及部署仍受到根本性瓶颈的制约。当前研究必须超越受控环境，解决人体生理的随机性以及边缘计算的严格限制。我们提出一份战略路线图，重点关注四个彼此独立的研究前沿。

A. 泛化差距：通过基础模型实现零样本迁移

生理信号固有的显著受试者间变异仍然是一个长期存在的障碍。舌形态和神经肌肉流形的解剖差异意味着严重的协变量偏移：即使语义意图 $P(Y|X)$ 保持不变，不同个体的输入分布$P(X)$ 也会发生巨大变化。为缓解这一问题，研究范式正转向跨模态基础模型。通过在海量无标注生物信号语料库和声学数据上进行预训练，这些模型能够学习一种通用的发音动作

潜在空间。正如大语言模型对自然语言处理产生的影响，我们设想借助神经语义对齐实现一种“即插即用”的体验：冻结的预训练编码器可将新用户的生物信号映射到共享嵌入空间中。这种零样本能力辅以元学习（例如 MAML）和领域对抗训练，旨在提取能够抵御电极位移和生理非平稳性的特征，并有望免除用户特定校准。

B. 设备端持续学习与硬件—算法协同设计

为了实现自然韵律，无声语音接口必须将闭环延迟控制在 ≤ 50 毫秒以内。这要求从基于云端的推理转向硬件感知边缘人工智能。然而，静态模型无法适应生物信号在数周使用过程中的时间漂移。未来研究必须优先关注设备端持续学习（OCL） [180], [181],，使接口能够在本地持续演进，同时避免灾难性遗忘。这一转变的核心是集成高能效的神经形态计算。在专用硬件上实现的脉冲神经网络（SNN）为信号处理提供了一条新路径，能够以事件驱动方式处理信号，模拟生物神经活动 [182]。通过将 OCL 与知识蒸馏、动态剪枝等算法策略相结合，基于 Transformer的解码器可以压缩至可穿戴设备形态，同时保留实时针对用户生理变化进行微调的能力。

C. 多模态大语言模型与潜在协同

单模态静默语音接口本质上受其截取点物理限制的约束，例如表面肌电对阻抗的敏感性，或超声易受声影影响。多模态融合的数学必要性在于最小化后验熵：$H(Y|X_{\text{combined}}) \leq \min(H(Y|X_i))$。一项重大突破是无声语音接口向多模态大语言模型（MLLM）的转变。通过对比目标将连续生物信号帧（例如表面肌电、超声舌成像、脑电图）与预训练大语言模型的嵌入空间对齐，系统可以利用模型庞大的语义先验来消解语音歧义。此外，应对“无声隆巴德效应” [183]——即缺乏听觉反馈所导致的发音漂移——需要闭环生物反馈。未来的多模态大语言模型应纳入触觉或视觉提示，以帮助用户在长期使用中于统一潜在流形内维持稳定的发音目标 [184]。

D. 认知自由的神经安全与伦理边界

随着无声语音接口向解码“内言语”迈进，它们突破了认知自由的关键边界。生理信号包含潜在身份标识，包括情绪状态和健康病理状况，从而使数据主权问题更为严峻。我们提出“神经安全”这一概念以界定生物数据所需的技术保障措施。技术防御必须整合联邦学习（FL）和差分隐私（DP），以确保原始信号保持本地化。此外，无声语音接口模型易受对抗扰动影响——恶意行为者可能注入“无声指令”——这要求开发稳健的、经过认证的防御层。迄今为止，既有的伦理框架必须防止未经同意的“神经监控”，并整合硬件级紧急关闭开关，以确保对私人思想的解码始终仅处于自主意志控制之下。

VII. 结论

无声语音接口已步入关键转折点，正从实验室原型迈向高性能、临床可行的基准系统。迄今，该领域由三大支柱共同定义：用于高保真感知的柔性生物电子器件、借助基于 Transformer 的架构实现的跨模态潜空间对齐，以及用于富有表现力、低延迟合成的生成式神经声码器。这一转变使无声语音接口从小众的恢复工具蜕变为普及型人机交互的基石，为传统依赖声学系统的结构性脆弱问题提供了明确的解决方案。

本综述综合的证据表明，大语言模型作为高层语义先验，已显著缓解“准确率瓶颈”。大语言模型充当神经纠错引擎，弥合了零散发音动作与连续自然语言之间的领域差距，实现了与最先进声学自动语音识别相当的性能。因此，研究前沿已从原始识别准确率转向长期稳健性与“无处不在而又隐形”的体验——将稳定的跨会话解码融入智能眼镜和耳戴式设备等日常可穿戴设备形态。

尽管取得了这些里程碑式的进展，“用户依赖悖论”仍阻碍着无声语音接口的普及。通过零样本迁移学习和自监督预训练克服解剖结构差异，仍是一项根本性挑战。此外，随着解码器逐渐触及意念表达的领域，“神经安全”的重要性日益凸显。要确保交流意图与私人思考之间的界限不受侵犯，就需要采用复杂的认知过滤器和安全的端侧处理架构，从而在日益互联的世界中维护心智自主权。

展望未来，我们预计这项技术将分阶段普及，先应用于康复医疗和隐蔽战术通信，随后再进入更广泛的消费市场。通过摆脱人类表达对声道声学限制的依赖，无声语音接口有望重新定义人机共生的边界。归根结底，无声语音接口技术能在沟通受限之处恢复人们的基本沟通权，并在对声音敏感的世界中增强数字自主性；它将从一种辅助性干预手段，逐渐演变为人类能力的无缝延伸。

VIII. 致谢

本研究获得国家科技重大专项（2023ZD0121101）和国防科技大学（ZZCX-ZZGC-01-04）的支持。

参考文献

[1] L. R. Rabiner，“隐马尔可夫模型及其在语音识别中的若干应用教程”，IEEE论文集，第77卷，第2期，第257–286页，2002年。[2] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，载于第31届神经信息处理系统国际会议论文集（NeurIPS 2017），2017年，第5998–6008页。[3] R. P. Lippmann，“机器与人类的语音识别”，《语音通信》，第22卷，第1期，第1–15页，1997年。[4] J. S. Edu、J. M. Such 和 G. Suarez-Tangil，“智能家居个人助理：安全与隐私综述”，ACM计算调查（CSUR），第53卷，第6期，第1–36页，2020年。[5] D. R. Beukelman 和 J. C. Light，《增强与替代沟通：为有复杂沟通需求的儿童和成人提供支持》，第5版。马里兰州巴尔的摩：Paul H. Brookes Publishing Co.，2020年。[6] B. Denby、T. Schultz、K. Honda、T. Hueber、J. M. Gilbert 和 J. S. Brumberg，“无声语音接口”，《语音通信》，第52卷，第4期，第270–287页，2010年。[7] Z. Ren、S. Pistrosch、B. Cos.kun、K. Scheck、A. Batliner、B. W. Schuller 和 T. Schultz，“无声副语言学导论”，arXiv预印本 arXiv:2508.18127，2025年。[8] J. A. Gonzalez-Lopez、A. Gomez-Alanis、J. M. Martin Donas、J. L. P´erez-C´ordoba 和 A. M. Gomez，“用于语音恢复的无声语音接口：综述”，IEEE Access，第8卷，第177 995–178 021页，2020年。[9] J. Freitas、A. Teixeira、M. S. Dias、S. Silva 等，《无声语音接口导论》。施普林格，2017年。[10] M. Zhang、Y. Yuan 和 S. Zhang，“观点章节：基于神经活动的无声语音接口——简要综述”，2025年。[11] F. R. Willett、E. M. Kunz、C. Fan、D. T. Avansino、G. H. Wilson、E. Y. Choi、S. N. Kamdar、M. F. Glasser、L. R. Hochberg、S. Druckmann 和 J. M. Henderson，“高性能语音神经假体”，Nature，第620卷，第7976期，第1031–1036页，2023年。[12] J. Zhou、D. Leong、Z. Cao、T. Do、S.-F. Liang 和 T.-P. Jung，“Mindspeak：用于无声语音的实时脑机接口系统”，载于第33届ACM多媒体国际会议论文集（MM ’25）。ACM，2025年，第1–3页。[13] A. Kapur、S. Kapur 和 P. Maes，“AlterEgo：个性化可穿戴无声语音接口”，载于第23届智能用户界面国际会议论文集（IUI ’18），2018年，第43–53页。[14] Y. Jin、Y. Gao、X. Xu、S. Choi、J. Li、F. Liu、Z. Li 和 Z. Jin，“Earcommand：在耳内‘听见’你的无声语音指令”，ACM交互式、移动、可穿戴与普适技术论文集，第6卷，第2期，2022年。[15] R. Zhang、W. Liu、K. L. Yang 等，“EchoSpeech：在轻便眼镜设备上实现连续无声语音识别”，载于2023年CHI计算系统人因会议论文集（CHI ’23），2023年。[16] N. Kimura、M. Kono 和 J. Rekimoto，“Sottovoce：基于超声成像和深度神经网络的无声语音交互”，载于2019年CHI人机交互计算系统会议论文集（CHI ’19），2019年，第1–11页。[17] J. A. Gonzalez、L. A. Cheah、J. M. Gilbert、J. Bai、S. R. Ell、P. D. Green 和 R. K. Moore，“基于永磁体发音动作描记和直接合成的无声语音系统”，《计算机语音与语言》，第39卷，第67–87页，2016年。[18] Q. Zhang、Y. Lan、K. Guo 和 D. Wang，“Lipwatch：利用声学传感在智能手表上实现无声语音识别”，ACM交互式、移动、可穿戴与普适技术论文集，第8卷，第2期，2024年。[19] D. Gaddy 和 D. Klein，“无声语音的数字发声”，载于2020年自然语言处理实证方法会议论文集（EMNLP），2020年，第5521–5530页。[20] T. Benster、G. Wilson、R. Elisha、F. R. Willett 和 S. Druckmann，“结合大语言模型增强识别的无声语音跨模态方法”，arXiv预印本 arXiv:2403.05583，2024年。[21] K. Su 和 L. Tian，“ 系统综述：基于脑电图的语音想象脑机接口解码与编码研究进展”，《PeerJ计算机科学》，第11卷，文章编号e2938，2025年。[22] Y. Min、J. Oh、J.-Y. Jeong 和 J.-W. Jeong，“探索公共VR/MR场景中通过无声语音接口进行情绪表达：自动化对用户体验的影响”，载于CHI人机交互计算系统会议扩展摘要集（CHI EA ’25），2025年

，第 1 页 –11.

[23] M. Eskenazi，“ 语音学与非语言学层面的语音与静默》，《语用学杂志》，第 43 卷，第 9 期，第 2286–2307 页，2011年。[24] B. Assouar、B. Liang、Y. Wu、Y. Li、J.-C. Cheng 和 Y. Jing，《声学超表面》，《自然评论：材料》，第 3 卷，472，2018 12，460 no. pp. – 。[25] M. Guo、Q. Fang、H. Zhang、W. Lu 和 J. Wei，《基于超声舌图像的无声语音识别跨模态元泛化》，《IEEE 音频、语音与语言处理汇刊》，第 34 卷，第 364–378 页，2025年。[26] A. Kaur、J. Singh、K. Boora 和 A. Garg，《基于 Convbi 集成深度学习的发声障碍患者肌电信号无声语音检测》，《生物医学信号处理与控制》，第 112 卷，文章编号 108386，2026年。[27] S. Hong、J. Yoo 和 S.-M. Park，《用于噪声环境下无声语音的人工智能驱动解码软性多轴应变映射界面》，《赛博格与仿生系统》。[28] M. E. Chowdhury、R. Sarmun、D. Bushnaq、M. Chabbouh、R. Aljindi、S. Pedersen、M. A. A. Faisal、N. Nahid、I. Hassan 和 C. Garcia，《使用脑电图数据进行阿拉伯语内言语识别的无声语音解码挑战总结》，《国际活动与行为计算杂志》，第 2025 卷，第 2 期，第 1–21 页，2025年。[29] G. Krishna、C. Tran、M. Carnahan 和 A. Tewfik，《使用脑电图进行连续无声语音识别》，arXiv 预印本 arXiv:2002.03851，2020年。[30] T. Kim、Y. Shin、K. Kang、K. Kim、G. Kim、Y. Byeon、H. Kim、Y. Gao、J. R. Lee、G. Son 等，《用于无声语音接口的超薄晶体硅应变计与深度学习算法》，《自然·通讯》，第 13 卷，第 1 期，文章编号 5815，2022年。[31] T. Y.-H. Chen、Y. Chen、P. Soederhaell、S. Agrawal 和 K. Shapo-valenko，《使用 Transformer 和基于 VAE 的数据增强解码脑电图语音感知》，arXiv 预印本 arXiv:2501.04359，2025年。[32] E. J. Wadkins，《用于 AlterEgo（一种无声语音接口）的连续无声语音识别系统》，博士论文，麻省理工学院，2019年。[33] S. Luo、M. Angrick、C. Coogan、D. N. Candrea、K. Wyse-Sookoo、A. Schippers、R. Ganji、G. W. Milsap、W. S. Anderson、C. R. Gordon 等，《用于设备控制的自定步调无声语音脑机接口》，medRxiv，第 2025–04 页，2025年。[34] S. D. Stavisky，《利用脑机接口恢复语音》，《生物医学工程年度评论》，第 27 卷，2025年。[35] N. S. Card、T. Singer-Clark、H. Peracha、C. Iacobacci、X. Hou、M. Wairagkar、Z. Fogg、E. Offenberg、L. R. Hochberg、D. M. Brandman 等，《用于语音与光标控制的皮层内脑机接口长期独立使用》，bioRxiv，第 2025–06 页，2025年。[36] J. S. Brumberg、A. Nieto-Castanon、P. R. Kennedy 和 F. H. Guenther，《用于语音通信的脑机接口》，《语音通信》，第 52 卷，第 4 期，第 367–379 页，2010年。[37] N. Fitriah、H. Zakaria 和 T. L. E. Rajab，《基于脑电图的无声语音接口及其挑战：综述》，《国际高级计算机科学与应用杂志》，第 13 卷，第 11 期，2022年。[38] K. Mohanchandra 和 S. Saha，《一种利用无声发音语音的通信范式：将脑信号转换为语音》，《增强人类研究》，第 1 卷，第 1 期，文章编号 3，2016年。[39] K. Tang、M. Zhao、H. Zhang、Y. Liu、Q. Xu 和 S. Wang，《用于长期无声语音监测与情绪识别的纺织品集成表面肌电传感器》，《自然·电子学》，第 8 卷，58，2025 1，45no. pp. – 。[40] A. K. Pattem、A. Illa、A. Afshan 和 P. K. Ghosh，《语音产生研究中电磁发音运动学记录的最佳传感器布置》，《计算机语音与语言》，第 47 卷，174，2018 157pp. – 。[41] P. Bonaventura、T. Piecuch、M. Oxley、J. Iida 和 J. Prahl，《通过电磁发音运动学评估戴假牙语音产生》，载于《2013 年第 35 届 IEEE 医学与生物学工程学会国际年会（EMBC）》。IEEE，2013年，第 4710–4713 页。[42] T. Rebernik、J. Jacobi、R. Jonkers、A. Noiray 和 M. Wieling，《使用电磁发音运动学的数据采集实践综述》，《实验室音系学》，第 12 卷，第 1 期，文章编号 6，2021年。[43] T. Srivastava、R. M. Winters、T. Gable、Y. T. Wang、T. LaScala 和 I. J. Tashev，《耳语式可穿戴设备：采用多模态方法，利用头戴式设备进行无声语音识别》，丛书：ICMI ’24。美国纽约：计算机协会，2024年，第 214–223 页。

[ ] 10 1145 3678957 3685720 在线。可访问：https://doi.org/ 。/ 。[44] J. Menezes、A. Mouras、A.-L. Fietkau、D. Kazzy 和 P. Birkholz，《利用雷达和腭位图无声语音接口进行音素多模态无声识别》

，” 载于 年度会议论文集

国际语音通信协会会议（Interspeech 2025），2025年，第5058–5062页。[45] D. Dash、E. Kittle、I. Gerrard、R. Csaky、G. Gonzalez、D. Taylor、J. P. Llinas、D. Labanowski、N. Deka和R. Yun，“基于可穿戴磁力计的无声语音识别，”bioRxiv，第2025 08页，2025年–。[46] R. A. Pizolato、F. S. d. F. Fernandes和M. B. D. Gavi˜ao，“颞下颌关节紊乱儿童的语音评估，”《应用口腔科学杂志》，第19卷，第493–499页，2011年。[47] L. Zhou、S. Chang、J. Luo、H. Wen、H. Zhu和L. Lu，“Baroauth：利用耳道变形在耳塞式耳机上实现说话用户认证，”载于2025年IEEE 第45届国际分布式计算系统会议（ICDCS）。IEEE，2025年，第111–121页。[48] Z. Wang、S. Tan、L. Zhang、Y. Ren、Z. Wang和J. Yang，“Eardynamic：一种基于耳道变形、使用入耳式可穿戴设备的连续用户认证，”ACM交互、移动、可穿戴与泛在技术论文集，第5卷，第1期，第1–27页，2021。[49] Y. Dai、M. Gao、R. Zhang、Y. Zhai、K. Cui和F. Xiao，“海报：面向商用现成耳机的惯性测量单元辅助语音增强，”载于第31届移动计算与网络国际年会论文集，2025年，第1383–1385页。[50] S. Mat Zin、S. Md Rasib、F. M. Suhaimi和M. Mariatti，“舌与硬腭接触检测技术：综述，”《生物医学工程在线》，第20卷，第1期，第17页，2021。[51] W. J. Hardcastle和F. Gibbon，“电腭图及其临床应用，”《仪器临床语音学》，第149–193页，1997年。[52] S. Preuß和P. Birkholz，“电光口腔造影的光学传感器校准。”载于国际语音通信协会年会论文集（Interspeech 2015），2015年，第618–622页。[53] S. Stone和P. Birkholz，“使用电光口腔造影的跨说话人无声语音命令词识别，”载于IEEE国际声学、语音与信号处理会议论文集（ICASSP 2020）。IEEE，2020年，第7849–7853页。[54] Y. Ji、L. Liu、H. Wang、Z. Liu、Z. Niu和B. Denby，“利用深度学习更新无声语音挑战基准，”《语音通信》，第98卷，第42–50页，2018。[55] B. Li、K. Xu、D. Feng、H. Mi、H. Wang和J. Zhu，“基于去噪卷积自编码器的B型超声舌图像特征提取，”载于IEEE国际声学、语音与信号处理会议论文集（ICASSP 2019）。IEEE，2019，第7130–7134页。[56] M. Stone，“从超声图像分析舌运动的指南，”Clinical Linguistics & Phonetics，第19卷，501，2005 6 7，455期-第–页。[57] Z. Xia、R. Yuan、Y. Cao、T. Sun、Y. Xiong和K. Xu，“机器学习技术应用于超声舌成像分析的系统综述，”美国声学学会期刊，第156卷，第3期，第1796–1819页，2024年。[58] K. Xu、T. G´abor Csap´o、P. Roussel和B. Denby，“超声舌像中带自动重新初始化的轮廓跟踪算法比较研究，”美国声学学会期刊，第139卷，第5期，第EL154–EL160页，2016年。[59] L. T´oth、A. H. Shandiz、G. Gosztolya和C. T. G´abor，“使用空间变换网络适配基于舌部超声的无声语音接口，”arXiv预印本arXiv:2305.19130，2023。[60] K. Xu、Y. Wu和Z. Gao，“使用序列卷积自编码器的基于超声的无声语音接口，”载于第27届ACM 多媒体国际会议论文集（MM ’19），2019，第2194–2195页。[61] A. Toutios和S. S. Narayanan，“面向语音科学与技术研究的声道实时磁共振成像进展，”《APSIPA信号与信息处理汇刊》，第5卷，第e6页，2016年。[62] S. Asadiabadi和E. Erzin，“使用深度时序回归网络在实时磁共振成像中跟踪声道轮廓，”IEEE/ACM 音频、语音与语言处理汇刊，第28卷，第3053–3064页，2020年。[63] K. Somandepalli、A. Toutios和S. S. Narayanan，“用于在实时磁共振图像中跟踪声道气体组织边界的语义边缘检测。”载于国际语音通信协会年会论文集（Interspeech 2017），2017年，第631–635页。[64] Y. Li、A. S. Hashim、Y. Lin、P. N. Nohuddin、K. Venkatachalam和A. Ahmadian，“面向元宇宙中逼真虚拟化身与唇读应用的基于人工智能的视觉语音识别，”《应用软计算》，第164卷，第111906页，2024年。

[65] Y. Chen、J. Yu、L. Kong 和 Y. Zhu，“侧信道声音感知方法综述”，IEEE物联网期刊，2024年。[66] Y. Igarashi、K. Futami 和 K. Murao，“静默语音眼镜接口：使用眼镜和配备红外距离传感器的耳戴式麦克风进行静默语音识别的方法”，Sensors，第24卷，第22期，第7368页，2024年。[67] J. Luo、J. Wang、N. Cheng、G. Jiang 和 J. Xiao，“基于声学传感的端到端无声语音识别”，载于2021年IEEE口语技术研讨会（SLT）。IEEE，2021年，第606–612页。[68] J. Zhou、D. Ding、Y. Li、Y. Lu、Y. Wang、Y. Zhang、Y.-C. Chen 和 G. Xue，“M2silent：通过共享空间中的多向扬声器实现多用户静默语音交互”，载于2025年CHI人机交互计算系统会议（CHI ’25）论文集，2025年，第1–19页。[69] C. Wagner、P. Schaffer、P. Amini Digehsara、M. B¨arhold、D. Plettemeier 和 P. Birkholz，“使用步进频率连续波雷达进行静默语音命令词识别”，《科学报告》，第12卷，第1期，第4192页，2022年。[70] S. Khan、L. Kallis、H. Mee、S. El Hadwe、D. Barone、P. Hutchinson 和 A. Kolias，“用于通信的侵入式脑机接口：范围综述”，《脑科学》，第15卷，第4期，第336页，2025年。[71] M. Inoue、M. Sato、K. Tomeoka、N. Nah、E. Hatakeyama、K. Arulku- maran、I. Horiguchi 和 S. Sasai，“基于异质电极配置的脑电图和肌电图无声语音解码系统”，arXiv预印本 arXiv:2506.13835，2025年。[72] N. Ramkumar 和 D. K. Renuka，“使用混合特征提取技术和集成堆叠分类器的新型基于脑机接口的无声语音识别：一种新型基于脑机接口的无声语音识别”，《科学与工业研究杂志》（JSIR），第82卷，第11期，第1165–1176页，2023年。[73] T. Lotey、S. Yadav 和 P. P. Roy，“使用深度学习技术的基于阿拉伯语脑电图的无声语音解码”，《国际活动与行为计算杂志》，第2025卷，第2期，第1–19页，2025年。[74] L. Xie、Y. Zhang、H. Yuan、M. Zhang、X. Zhang、C. Zheng、Y. Yan 和 E. Yin，“结合面部肌电图的神经中文无声语音识别”，《语音通信》，第171卷，第103230页，2025年。[75] A. T. Chowdhury、M. Newaz、P. Saha、M. N. AbuHaweeleh、S. Mohsen、D. Bushnaq、M. Chabbouh、R. Aljindi、S. Pedersen 和 M. E. Chowd- hury，“无声语音解码：从数据、方法和框架角度看机器学习”，《神经计算与应用》，第37卷，第10期，第6995–7013页，2025年。[76] G. S. Meltzner、J. T. Heaton、Y. Deng、G. De Luca、S. H. Roy 和 J. C. Kline，“无声语音识别：喉切除术患者的替代通信设备”，IEEE/ACM 音频、语音与语言处理汇刊，第25卷，第12期，第2386–2398页，2017年。[77] M. Janke 和 L. Diener，“肌电到语音：从面部肌电信号直接生成语音”，IEEE/ACM 音频、语音与语言处理汇刊，第25卷，第12期，第2375–2385页，2017年。[78] C. Dai、S. Yan、M. S. Tong 和 J. Chen，“一种基于表面肌电信号的新型无声语音识别系统设计”，载于2025年光子学与电磁学研究研讨会—春季会议（PIERS-Spring）。IEEE，2025年，第1–6页。[79] J. Freitas、A. Teixeira、J. Dias、A. Ferreira 和 M. Figueiredo，“通过特征选择增强多模态静默语音接口”，载于国际语音通信协会年会论文集（Interspeech 2014）。语音与通信协会，2014年。[80] A. Abdullah 和 K. Chemmangat，“ 一种计算高效的基于表面肌电的无声语音接口：采用通道削减和基于决策树的分类”，Procedia Computer Science，第171卷，第120–129页，2020年。[81] I. Ibrahimov、Z. Csaba 和 G. Gosztolya，“基于Conformer的超声到语音转换”，arXiv预印本 arXiv:2506 03831，2025年。。[82] R.-C. Zheng、Y. Ai 和 Z.-H. Ling，“融合超声舌图像进行视听语音增强”，IEEE/ACM 音频、语音与语言处理汇刊，第32卷，1444，2024年 1430页。– 。[83] S. Stone，《一种使用电光口腔造影的无声语音接口》。BoD–Books on Demand，2021年。[84] T. Hueber、G. Aversano、G. Chollet、B. Denby、G. Dreyfus、Y. Ous- sar、P. Roussel 和 M. Stone，“用于基于超声的无声语音接口的Eigentongue特征提取”，见论文集

2007年 IEEE国际声学、语音与信号处理会议（ICASSP），2007年，第 1–1245 页。[85] K. You、B. Liu、K. Xu、Y. Xiong、Q. Xu、M. Feng、T. G. Csap´o 和 B. Zhu，“基于原始超声的音素片段分类：掩码建模方法”，载于 2023 IEEE国际声学、语音与信号处理会议（ICASSP）论文集，2023年，第 1–5 页。[86] C. Wu、S. Chen、G. Sheng、P. Roussel 和 B. Denby，“使用三维卷积神经网络预测未标注超声视频中的舌运动”，载于 2018 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2018年，第 5764–5768 页。[87] J. Cai、B. Denby、P. Roussel、G. Dreyfus 和 L. Crevier-Buchman，“采用语言模型的轻量级超声无声语音接口的识别与实时性能”，载于国际语音通信协会第 12 届年会论文集（Interspeech），2011年。[88] S. Stone 和 P. Birkholz，“使用电光口腔造影进行无声语音命令词识别”，载于国际语音通信协会年会论文集（Interspeech 2016），2016年，第 2350–2351 页。[89] N. Kimura、Z. Su 和 T. Saeki，“面向无声语音挑战的端到端深度学习语音识别模型”，载于国际语音通信协会年会论文集（Interspeech 2020），2020年，第 1025–1026 页。[90] G. Gosztolya、 ´A. Pint´er、L. T´oth、T. Gr´osz、A. Mark´o 和 T. G. Csap´o，“基于自编码器的发音动作到声学映射：用于超声无声语音接口”，载于 2019年国际神经网络联合会议（IJCNN）。IEEE，2019年，第 1–8 页。[91] M. Kim、B. Cao、T. Mau 和 J. Wang，“利用肉体表面发音动作和 LSTM 神经网络进行说话人无关的无声语音识别”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 25 卷，第 12 期，第 2323–2336 页，2017年。[92] J. Wang 和 S. Hahm，“采用跨说话人发音动作归一化和说话人自适应训练的说话人无关无声语音识别”，载于国际语音通信协会年会论文集（Interspeech 2015），2015年，第 2415–2419 页。[93] W. Lee、J. J. Seong、B. Ozlu、B. S. Shim、A. Marakhimov 和 S. Lee，“生物信号传感器与基于深度学习的语音识别：综述”，Sensors，第 21 卷，第 4 期，第 1399 页，2021年。[94] Z. Su、S. Fang 和 J. Rekimoto，“基于多模态无声语音、由词首字母条件化的大语言模型文本输入”，载于 ACM 第 7 届对话式用户界面会议论文集，2025年，第 1–14 页。[95] T. Liu、M. Zhang、Z. Li、H. Dou、W. Zhang、J. Yang、P. Wu、D. Li 和 X. Mu，“机器学习辅助的可穿戴语音识别与交互传感系统”，《自然·通讯》，第 16 卷，第 1 期，第 2363 页，2025年。[96] S. Zeng、H. Wan、S. Shi 和 W. Wang，“mSilent：利用商用毫米波雷达实现通用语料库无声语音识别”，ACM《交互式、移动、可穿戴与普适技术论文集》，第 7 卷，第 1 期，第 1–28 页，2023年。[97] X. Sun、J. Xiong、C. Feng、H. Li、Y. Wu、D. Fang 和 X. Chen，“Earssr：通过耳机实现无声语音识别”，《IEEE移动计算汇刊》，第 23 卷，第 8 期，第 8493–8507 页，2024年。[98] M. Fukumoto，“Silentvoice：通过吸气语音实现不易察觉的语音输入”，载于 ACM 第 31 届用户界面软件与技术年度研讨会论文集（UIST ’18），2018年，第 237–246 页。[99] J. Taylor 和 K. Richmond，“端到端语音合成中的发音学习分析”，载于国际语音通信协会第 20 届年会：语音与语言的交汇点。国际语音通信协会，2019年，第 2070–2074 页。[100] R. Prabhavalkar、T. Hori、T. N. Sainath、R. Schl¨uter 和 S. Watanabe，“端到端语音识别：综述”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 32 卷，351，2023 325页。[101] R. N. Khushaba、A. Phinyomark、A. H. Al-Timemy 和 E. Scheme，“结合注意力机制的递归多信号时序融合改进了肌电图特征提取”，《IEEE人工智能汇刊》，第 1 卷，第 2 期，第 139–150 页，2020 年。[102] T. Schultz 、 M. Wand、T. Hueber、D. J. Krusienski、C. Herff 和 J. S. Brumberg，“基于生物信号的口语交流：综述”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 25 卷，第 12 期，第 2257–2271 页，2017年。[103] Y. Lim、P. Kumar 和 K. S. Nayak，“0.55 T 实时磁共振成像下的语音产生”，Magnetic Resonance in Medicine，第 91 卷，第 1 期，第 337–343 页，2024年。

[104] Y. Lan、Dawson 和 Z. Zhou，《使用超声成像量化舌形状和复杂度的方法》，Clinical Linguistics & Phonetics，第30卷，第3–5期，第328–344页，2016年。[105] S. G. Lingala、Y. Zhu、Y.-C. Kim、A. Toutios、S. Narayanan和K. S. Nayak，《用于研究动态声道塑形的快速灵活磁共振成像系统》，Magnetic Resonance in Medicine，第77卷，第1期，第112–125页，2017年。[106] C. Bregler和Y. Konig，《“特征唇”：用于稳健语音识别》，载于《ICASSP’94论文集》。IEEE国际声学、语音与信号处理会议，第2卷。IEEE，1994年，第II–669页。[107] J. Berry、D. Archangeli和I. Fasel，《超声图像中舌部手势的自动分类》，载于第12届实验音系学会议（LabPhon）论文集，2010年。[108] D. Jenson、A. L. Bowers、A. W. Harkrider、D. Thornton、M. Cuellar和T. Saltuklaroglu，《语音感知与产生中感觉运动整合的时间动态：脑电图数据的独立成分分析》，Frontiers in Psychology，第5卷，第656页，2014年。[109] B. Vachhani、C. Bhat、B. Das和S. K. Kopparapu，《基于深度自编码器的语音特征与构音障碍语音识别改进》，载于Interspeech，2017年，第1854–1858页。[110] Y. Qiu、W. Zhou、N. Yu和P. Du，《基于去噪稀疏自编码器的发作期脑电分类》，IEEE神经系统与康复工程汇刊，第26卷，第9期，第1717–1726页，2018年。[111] M. Angrick、C. Herff、E. Mugler、M. C. Tate、M. W. Slutzky、D. J. Krusienski和T. Schultz，《使用密集连接的三维卷积神经网络从皮层脑电图合成语音》，Journal of Neural Engineering，第16卷，第3期，第036019页，2019年。[112] Y.-H. Lai、F. Chen、S.-S. Wang、X. Lu、Y. Tsao和C.-H. Lee，《采用深度去噪自编码器提高人工耳蜗模拟中声码器处理语音的可懂度》，IEEE生物医学工程汇刊，第64卷，第7期，第1568–1578页，2016年。[113] H. Kabakoff、D. Harel、M. Tiede、D. Whalen和T. McAllister，《将超声舌形复杂度测量扩展至语音发育与障碍》，Journal of Speech, Language, and Hearing Research，第64卷，第7期，第2557–2574页，2021年。[114] R. Wilhelms-Tricarico，《语音产生的生理建模：软组织发音器官的建模方法》，美国声学学会期刊，第97卷，第5期，第3085–3098页，1995年。[115] E. Sifakis、A. Selle、A. Robinson-Mosher和R. Fedkiw，《使用基于物理的面部肌肉模型模拟语音》，载于2006年ACM SIGGRAPH/Eurographics计算机动画研讨会论文集，2006年，第261–270页。[116] Y. Yang、R. Su、S. Zhao、M. L. Ng、N. Yan和L. Wang，《面向统一的语音到超声舌成像合成扩散模型》，Information Fusion，第103896页，2025年。[117] K. Xu、Y. Yang、C. Leboullenger、P. Roussel和B. Denby，《基于超声图像序列的轮廓式三维舌运动可视化》，载于2016年IEEE国际声学、语音与信号处理会议（ICASSP）论文集。IEEE，2016年，第5380–5384页。[118] K. Xu、J. Zhao、B. Zhu和C. Zhao，《利用序列到序列学习从唇部图像预测超声舌图像》，美国声学学会期刊，第147卷，第6期，第EL441–EL446页，2020年。[119] A. Wrench和J. Balch-Tomes，《超越边缘：使用DeepLabCut从超声和相机图像中进行语音发音器官的无标记姿态估计》，Sensors，第22卷，第3期，第1133页，2022年。[120] J. Sun、T. Kitamura和R. Hayashi，《使用DeepLabCut从超声和实时磁共振成像数据中提取语音器官轮廓》，Acoustical Science and Technology，第46卷，第4期，第338–344页，2025年。[121] C. Juan、O. Ramos和D. Amaya，《无声语音的当前技术》，2016年。[122] G. K. Anumanchipalli、J. Chartier和E. F. Chang，《从口语句子的神经解码合成语音》，Nature，第568卷，第7753期，第493–498页，2019年。[123] K. Scheck、T. Dombeck、Z. Ren、P. Wu、M. Wand和T. Schultz，《DiffMV-ETS：使用说话人无关的语音训练目标，基于扩散的多说话人肌电图到语音转换》，载于国际语音通信协会年度会议（Interspeech 2025）论文集，2025年，第5573–5577页。[124] H. Liu、W. Dong、Y. Li、F. Li、J. Geng、M. Zhu、T. Chen、H. Zhang、L. Sun和C. Lee，《用于失声患者的新型人机接口：表皮式肌电纹身贴片》，Microsystems & Nanoengineering，第6卷，第1期，第16页，2020年。[125] D. Gaddy和D. Klein，《改进的无声语音发声模型》，arXiv预印本，arXiv:2106.01933

[126] J. A. Gonzalez、L. A. Cheah、A. M. Gomez、P. D. Green、J. M. Gilbert、S. R. Ell、R. K. Moore 和 E. Holdsworth，“通过机器学习从发音传感器数据重建直接语音”，IEEE/ACM 音频、语音与语言处理汇刊，第25卷，第12期，第2362–2374页，2017年。[127] T. G. Csap´o、C. Zaink´o、L. T´oth、G. Gosztolya 和 A. Mark´o，“结合WaveGlow语音合成的基于超声的发音到声学映射”，arXiv预印本 arXiv:2008.03152，2020年。[128] R.-C. Zheng、Y. Ai 和 Z.-H. Ling，“通过扩散模型和文本引导的伪目标生成，从无声唇舌发音重建语音”，载于第32届ACM国际多媒体会议论文集，2024年，第6559–6568页。[129] T. Hueber 和 G. Bailly，“使用全协方差隐马尔可夫模型将无声发音统计转换为可听语音”，《计算机语音与语言》，第36卷，第274–293页，2016年。[130] L. J.-T. Yu、R. Zhao、S. Ji、E. C. H. Ngai 和 C. Wu，“Uspeech：通过跨模态合成以最小人工 effort 实现超声增强语音”，ACM交互、移动、可穿戴与泛在技术论文集，第9卷，第2期，2025年。[131] T. Hirahara、M. Otani、S. Shimizu、T. Toda、K. Nakamura、Y. Nakajima 和 K. Shikano，“利用体传导声道共振信号进行无声语音增强”，《语音通信》，第52卷，第4期，第301–313页，2010年。[132] N. Kimura、T. Gemicioglu、J. Womack、R. Li、Y. Zhao、A. Bedri、Z. Su、A. Olwal、J. Rekimoto 和 T. Starner，“Silentspeller：利用电腭图实现面向移动设备、免手持的无声语音文本输入”，载于2022年CHI计算系统人因会议论文集（CHI ’22），2022年，第1–19页。[133] N. Sivasubramaniam，“从无声信号到自然语言：双阶段Transformer-大语言模型方法”，arXiv预印本 arXiv:2509 04507，2025年。。[134] P. Mohapatra、A. Pandey、X. Zhang 和 Q. Zhu，“大语言模型能理解清音语音吗？探索利用大语言模型进行肌电信号转文本转换”，arXiv预印本 arXiv:2506.00304，2025年。[135] J. S. Kang、K. S. Moon、S. Q. Lee、N. Satterlee 和 X. Zuo，“使用肌电图和压电传感器的可穿戴无声文本输入系统”，Sensors，第25卷，第8期，第2624页，2025年。[136] T. Hueber、P. Badin、G. Bailly、A. B. Youssef、F. Elisei、B. Denby 和 G. Chollet，“发音与声学数据之间的统计映射：在无声语音接口和可视化发音反馈中的应用”，载于P3S第一届表演语音与歌唱合成国际研讨会 *[P3S]*, 2011。[137] Z. Su、S. Fang 和 J. Rekimoto，“Liplearner：移动设备上的可定制无声语音交互”，载于2023年CHI计算系统人因会议论文集（CHI ’23），2023年，第1–21页。[138] C. Tang、M. Xu、W. Yi、Z. Zhang、E. Occhipinti、C. Dong、D. Raven-scroft、S.-M. Jung、S. Lee、S. Gao 等，“超灵敏纺织品应变传感器以高机器学习效率重新定义可穿戴无声语音接口”，《npj柔性电子学》，第8卷，第1期，第27页，2024年。[139] S. Hahm、J. Wang 和 J. Friedman，“ 使用深度神经网络从发音动作中进行无声语音识别 ”，载于国际语音科学会议，2015年。[140] X. Wang、Z. Su、J. Rekimoto 和 Y. Zhang，“观察你的嘴：结合深度传感的无声语音识别”，载于2024年CHI计算系统人因会议论文集（CHI ’24），2024年。[141] T. Gr´osz、G. Gosztolya、L. T´oth、T. G. Csap´o 和 A. Mark´o，“基于深度神经网络的超声无声语音接口的基频估计”，载于IEEE国际声学、语音与信号处理会议论文集（ICASSP 2018）。IEEE，2018年，第291–295页。[142] R.-C. Zheng、Y. Ai 和 Z.-H. Ling，“通过伪目标生成和领域对抗训练从无声舌唇发音重建语音”，载于IEEE国际声学、语音与信号处理会议论文集（ICASSP 2023）。IEEE，2023年，第1–5页。[143] I. Hernaez、J. A. Gonz´alez-L´opez、E. Navas、J. L. P´erez C´ordoba、I. Saratxaga、G. Olivares、J. S´anchez de la Fuente、A. Gald´on、V. Garc´ıa Romillo、M. Gonz´alez-Atienza 等，“利用无声语音接口进行语音恢复（RESSINT）”，载于 IberSPEECH 2021论文集，2021年，第130–134页。[144] F. Meier、G. Spacone、S. Frey、L. Benini 和 A. Cossettini，“基于可穿戴全干式肌电颈带的并行超低功耗无声语音接口”，arXiv预印本 arXiv:2509.21964，2025年。[145] N. Kimura、Z. Su、T. Saeki 和 J. Rekimoto，“SSR7000：用于端到端无声语音识别的同步超声舌成像语料库”，载于第十三届语言资源与评测会议论文集，2022年，第6866–6873页。

[146] R. Beeson 和 K. Richmond，《利用舌部超声与唇部视频估计发音器官位置的无声语音识别》，载于国际语音通信协会年会论文集（Interspeech 2023）。ISCA，2023年，第1149–1153页。[147] H. Wang、P. Roussel 和 B. Denby，《无声语音接口的舌部动态表征学习》，IEICE TRANSACTIONS on Information and Systems，第104卷第12期，第2209–2217页，2021年。[148] Y. Yang、R. Su、S. Zhao、J. Wei、M. L. Ng、N. Yan 和 L. Wang，《用于普通话语音的音频与超声同步舌部运动数据库》，Scientific Data，第12卷第1期，第607页，2025年。[149] D. Zhou、Y. Zhang、J. Wu、X. Zhang、L. Xie 和 E. Yin，《AVE Speech：融合音频、视觉和肌电信号的综合多模态语音识别数据集》，IEEE Transactions on Human-Machine Systems，2025年。[150] J. Zhou、Z. Cao、Y. Duan、C. Barkley、D. Leong、X. Jiang、Q.-T. Nguyen、Z. Zhao、T. Do、Y.-C. Chang 等，《面向主动脑机接口的大型脑语言模型预训练：无声语音》，载于第33届 ACM 多媒体国际会议（MM ’25）论文集。ACM，2025年，第1–10页。[151] N. Shah、S. Karande 和 V. Gandhi，《利用新方法和 MultiNAM 数据集推进神经声学测量到语音的转换》，载于 IEEE国际声学、语音与信号处理会议（ICASSP 2025）论文集。IEEE，2025年，第1–5页。[152] M. Wand、M. Janke 和 T. Schultz，《用于肌电语音处理的 EMG-UKA 语料库》，载于国际语音通信协会年会论文集（Interspeech 2014），2014年，第1593–1597页。[153] V. Sivakumar、J. Seely、A. Du、S. Bittner、A. Berenzweig、A. Bolarinwa、A. Gramfort 和 M. Mandel，《EMG2QWERTY：用于表面肌电盲打的大规模数据集及基准》，神经信息处理系统进展，第37卷，91 389，2024年 91 373页。– 。[154] B. Denby 和 M. Stone，《从舌部超声图像合成语音》，ICASSP 论文集，加拿大蒙特利尔，2004年。´M. S. Ribeiro、J. Sanger、J.-X. Zhang、A. Eshky、A. Wrench、K. Richmond 和 S. Renals，《TAL：包含超声舌成像、音频和唇部视频的同步多说话人语料库》，载于2021 IEEE 口语技术研讨会（SLT）。IEEE，2021年，第1109–1116页。[155] A. Eshky、M. S. Ribeiro、J. Cleland、K. Richmond、Z. Roxburgh、J. Scobbie 和 A. Wrench，《UltraSuite：来自儿童言语治疗课程的超声与声学数据存储库》，arXiv 预印本 arXiv:1907.00835，2019年。[156] T. Afouras、J. S. Chung 和 A. Zisserman，《LRS3-TED：用于视觉语音识别的大规模数据集》，arXiv 预印本 arXiv:1809.00496，2018年。[157] S. Yang、Y. Zhang、D. Feng、M. Yang、C. Wang、J. Xiao、K. Long、S. Shan 和 X. Chen，《LRW-1000：一个自然分布的大规模野外唇读基准》，载于2019年第14届 IEEE 自动人脸与手势识别国际会议（FG 2019）。IEEE，2019年，第1–8页。[158] N. Nieto、V. Peterson、H. L. Rufiner、J. E. Kamienkowski 和 R. Spies，《大声思考：一个用于内言语识别的开放获取基于脑电图的脑机接口数据集》，Scientific Data，第9卷第1期，第52页，2022年。[159] J. R. Brennan 和 J. T. Hale，《层级结构引导自然情境聆听过程中的快速语言预测》，PLOS ONE，第14卷第1期，第e0207741页，2019年。[160] B. Accou、L. Bollens、M. Gillis、W. Verheijen、H. Van Hamme 和 T. Francart，《SPARRKULEE：来自鲁汶大学的语音诱发听觉反应存储库，包含85名参与者的脑电图数据》，Data，第9卷第8期，第94页，2024年。[161] A. Wrench，《多通道发音语音数据库及其在自动语音识别中的应用》，载于第5届语音产生研讨会：模型与数据论文集，2000年。[162] S. Narayanan、A. Toutios、V. Ramanarayanan、A. Lammert、J. Kim、S. Lee、K. Nayak、Y.-C. Kim、Y. Zhu、L. Goldstein 等，《用于语音产生研究的实时磁共振成像与电磁发音运动学数据库（TC）》，《美国声学学会期刊》，第136卷第3期，第1307–1311页，2014年。[163] K. Richmond、P. Hoole 和 S. King，《发布 MNGU0 发音语料库的电磁发音运动学（第1天）子集》，载于国际语音通信协会年会论文集（Interspeech 2011）。国际语音通信协会，2011年，第1505–1508页。[164] Y. Kim、Y. Song 和 Y. Chung，《TAPS：用于基于深度学习的语音增强的喉部与声学配对语音数据集》，arXiv 预印本 arXiv:2502.11478, 2025年。

[166] H. T. Gowda 和 L. M. Miller，“非侵入式肌电语音神经假体：几何视角”，*arXiv*预印本 *arXiv:2502.05762*，2025年。[167] M. S. Ribeiro、A. Eshky、K. Richmond 和 S. Renals，“基于超声和视频的静音与常规多说话人语音识别”，*arXiv*预印本 *arXiv:2103.00333*，2021年。[168] V. Ramanarayanan、S. Tilsen、M. Proctor、J. T¨oger、L. Goldstein、K. S. Nayak 和 S. Narayanan，“语音产生的实时磁共振成像分析”，《计算机语音与语言》，第52卷，第1–22页，2018年。[169] B. Cao、S. Ravi、N. Sebkhi、A. Bhavsar、O. T. Inan、W. Xu 和 J. Wang，“Magtrack：用于无声语音接口的可穿戴舌运动跟踪系统”，*Journal of Speech, Language, and Hearing Research*，第66卷，第8S期，第3206–3221页，2023年。[170] E. M. Juanpere 和 T. G. Csap´o，“基于超声的无声语音接口：卷积神经网络与循环神经网络的应用”，《声学学报》，第105卷，第4期，第587–590页，2019年。[171] A. Finkelstein，“借助直接语音脑机接口打破沉默：在关于缓解算法偏见的伦理建议中，将沟通障碍置于核心”，*GRACE*：全球人工智能社区伦理评论，第3卷，第1期，2025年。[172] A. Garcia，“通过语音转文本迁移学习实现基于肌电的无声语音转录”，硕士论文，加州州立理工大学，2024年。[173] P. Dai、M. S. Al-Radhi 和 T. G. Csap´o，“基频估计算法对基于超声的无声语音接口的影响”，载于*2021*年国际语音技术与人机对话会议（*SpeD*）。IEEE，2021年，第47–51页。[174] L. Pandey、K. Hasan 和 A. S. Arif，“在私人和公共场合使用语音和无声语音输入方法的可接受性”，载于*2021*年*CHI*人机交互计算系统会议论文集（*CHI '21*），2021年，第1–13页。[175] X. Dong、L. Xu、L. He、Z. Han、K. Christofferson、Y. Chen、A. Taya、Y. Nishiyama 和 K. Sezaki，“海报：利用多任务学习识别耳内隐藏式私密密钥，实现可靠的无声语音接口”，载于*2025*年*ACM*普适与泛在计算国际联合会议伴随会议（*UbiComp Companion '25*）。ACM，2025年，第1–5页。[176] S. Lee 和 J. Yu，“基于脉冲无线超宽带雷达的非接触式无声语音识别：采用注意力增强型时序卷积网络”，载于*2025*年*IEEE/IEIE*国际消费电子-亚洲会议（*ICCE-Asia*）。IEEE，2025年，第1–5页。[177] P. Dong、Y. Li、S. Chen、J. T. Grafstein、I. Khan 和 S. Yao，“通过柔性磁性皮肤和机器学习，从发音动作中解码无声语音指令”，《材料视野》，第10卷，第12期，第5607–5620页，2023年。[178] S. Xu、J.-X. Yu、H. Guo、S. Tian、Y. Long、J. Yang 和 L. Zhang，“两性离子水凝胶中的力诱导离子生成：一种灵敏的无声语音传感器”，《自然·通讯》，第14卷，219，2023年1，第.页。[179] C. Tang、J. Mallah、D. Kazieczko、W. Yi、T. R. Kandukuri、E. Oc-chipinti、B. Mishra、S. Mehta 和 L. G. Occhipinti，“集成于耳机中的多通道纺织肌电传感器无线无声语音接口”，《*IEEE*仪器与测量汇刊》，2025年。[180] C. E. Taylor、S. M. Patil 和 C. Dovrolis，“遗忘之前，先有学习：在线无监督持续学习中的表征学习挑战”，《机器学习研究汇刊》。[181] F. Lyu、G. Cheng、D. Liu、L. Zhao、Z. Zhang、F. Hu 和 L. Wang，“通过双边距对比重放缓解在线持续学习中的灾难性遗忘”，*IEEE*电路与系统视频技术汇刊，2025年。[182] M. Bouvier、A. Valentian、T. Mesquida、F. Rummens、M. Reyboz、E. Vianello 和 E. Beigne，“脉冲神经网络的硬件实现与挑战：综述”，《*ACM*计算系统新兴技术期刊》（*JETC*），第15卷，35，2019年2，第1期，页码–。[183] J. Luo、S. R. Hage 和 C. F. Moss，“隆巴德效应：从声学到神经机制”，《神经科学趋势》，第41卷，第12期，第938–949页，2018年。[184] G. Huang、D. H. K. Tsang 和 L. Liu，“伸出援手：通过多模态大语言模型驱动的手部建模，实现半免训练提示语识别，促进无障碍沟通”，*arXiv*预印本 *arXiv:2503.21785*，2025年。



Kele Xu （IEEE高级会员）于2017年在法国巴黎第六大学获得博士学位。现任中国国防科技大学计算机科学学院副教授。他在ICML、CVPR、NeurIPS、ICLR、AAAI、IJCAI、ASE和ACM多媒体等期刊和会议上发表或合作发表了100余篇论文。其研究方向包括音频信号处理、机器学习和智能软件系统。他还担任《IEEE视频技术电路与系统汇刊》的副编辑，以及《科学合作期刊：仿生与仿生系统》的客座编辑。

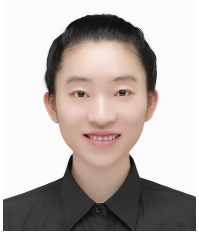
统汇刊的副编辑，以及《科学合作期刊：仿生与仿生系统》的客座编辑。



Yifan Wang 分别于2020年和2025年在中国青岛的中国海洋大学获得工学学士学位和博士学位。她目前是中国长沙国防科技大学计算机科学学院的博士后研究人员。她的研究兴趣包括音频信号处理和识别。目前，她的研究重点是学习算法、主动学习和神经网络设计。



Qisheng Xu 是国防科技大学计算机科学与技术学院的博士研究生。他于2024年获得国防科技大学理学硕士学位，并于2021年获得武汉大学理学学士学位。他的研究兴趣包括音频信号处理、水下声学目标识别和持续学习算法。



Wuyang Chen 于2025年在中国长沙的国防科技大学获得计算机科学与技术博士学位。她目前是中国长沙湖南师范大学信息科学与工程学院的讲师。她的研究兴趣包括多模态学习、语音-人脸关联及相关领域。



Yutao Dou 于2020年获得澳大利亚堪培拉大学（位于澳大利亚首都领地堪培拉）的软件工程工学学士学位；于2022年获得悉尼大学哲学硕士学位；目前就读于中国长沙湖南大学计算机科学与电子工程学院，攻读博士学位。他的研究兴趣主要包括分布式计算、生物信息学和人工智能。



Ming Feng 于2025年获得中国上海同济大学博士学位；目前在中国长沙国防科技大学担任软件工程博士后研究人员。他的研究兴趣主要包括语音处理、医学图像处理、弱监督学习、手写识别及相关人工智能技术。



Cheng Yang 是国防科技大学（NUDT）的副教授。他于2018年在国防科技大学获得计算机科学博士学位。他的研究兴趣包括音频信号处理与识别。目前的研究重点是学习算法、增量学习和神经网络设计。



Huaimin Wang 是中国科学院院士、国防科技大学教授。他于1992年在国防科技大学获得计算机科学博士学位。他还获得了国家杰出青年科学基金资助。Wang博士已在知名国际会议和期刊上发表了100多篇研究论文。目前，他的研究兴趣集中在中间件、软件代理和可信计算。