语言学家应学会爱上基于语音的深度学习模型

Marianne de Heer Kloots[1], Paul Boersma[2], Willem Zuidema[1] 阿姆斯特丹大学逻辑、语言与计算研究所 [2] 阿姆斯特丹大学阿姆斯特丹语言与传播中心

通信地址：m.l.s.deheerkloots@uva.nl

关于 Futrell, R. 和 Mahowald, K.（即将出版）的评论。语言学如何学会不再担忧并爱上语言模型。《行为与脑科学》。http://doi.org/10.1017/S0140525X2510112X

摘要

Futrell 和 Mahowald 提出了一个有用的框架，将技术导向的深度学习系统与解释导向的语

正文

要对人类口语进行综合研究，语言学家会考察说话者如何将交际意图一路转化为发音动作

文本的问题。在语言理解方面，大语言模型处理的是文本；文本已经将连续的听觉语流离散化为类似词、语素、音节和／或音系音段（元音、辅音）的单位，具体取决于语言。反过来，口语中至关重要的音系特征，如韵律和语调，通常不会体现在文本中。因此，文本和语音中的歧义并不相同：英语文本能区分*sun*和*son*，英语语音却不能；而英语语音能消除你喜欢咖啡还是茶？这一问句中是非选择与选择解读之间的歧义，文本却不能。总的来说，正如认知科学家和语言技术专家此前所指出的，基于文本的模型最终解决的问题，与人类口语使用者面对的根本不同（Dupoux，2018；Chrupała，2023）。

自然语言的结构在极大程度上反映的是语音（或手语）的特性，而非文本的特性。仅仅通过声学标记或独立的语音识别／合成组件为基于文本的大语言模型增添语音能力（Arora 等，2025），并不能解决语言建模中文本瓶颈所造成的深层问题。因此，研究语言结构建模的研究者不应满足于大语言模型，而应学会喜爱那些旨在捕捉人类口语更自然形式的模型：语音信号本身。

语音模型中的语言结构。如今，许多语音技术应用都采用自监督语音基础模型，即基于未标注语音录音训练、用于表征音频信号的深度学习架构。与研究基于文本的大语言模型可解释性（如 F&M 所述）的工作相似，语言可解释性研究探讨语音领域中的基于语音的模型在多大程度上学会捕捉构成口语的高层模式。在方法上，这些研究大多使用表征探针来考察哪些语言

信息被表征在语音模型的内部状态中；对于音位（Alishahi, Barking, & Chrupała, 2016; Martin et al., 2023）和词（Pasad et al., 2024）等语言单位的编码，以及形态音系模式（Gauthier et al., 2025）和超音段模式（Shen et al., 2024）的编码，研究均取得了颇有启发性的结果；其他研究则采用行为（最小对立体）测试，例如探究模型能否区分词与伪词（Lavechin et al., 2023），或韵律自然与不自然的停顿（de Seyssel et al., 2023）。借助这些方法，语音模型可以充当“心理语言学参与者”，模拟人类言语感知实验及其实验结果，例如有关感知相似性（Millet & Dunbar, 2022）和语音范畴化（de Heer Kloots & Zuidema, 2024）的结果。

语言学家已经可以从这些发现中得出一个重要的理论见解：纯语音模型确实能够学习与语言结构相关的模式，

来自领域一般性感知的归纳偏置。 进一步研究深度学习模型从音频中学到了什么，或许能够帮

双向处理的归纳偏置。 许多认知基础神经建模都具有一个当前技术语音或文本机器所不具备的

在语音学和音系学中，具有双向连接权重的神经网络模型可以直接预测音位库中听觉分散度的

小型神经网络模型的一个缺点是，它们只能处理玩具问题，而非完整语言。希望玩具模型的证

核心要旨。我们认同，为技术应用开发的神经网络模型能够帮助我们深入认识人类语言与认知。

参考文献

Alishahi, A., Barking, M., & Chrupała, G. (2017). 具身语音循环神经模型中的音系编码。载于《第

Wu, H

发展全 550/arXiv.2504.08528 Aslin, R.,

& Piso

Kavan 《儿童音系：第2卷，感知》（第67–96页）。纽约：学

发展。载于 John P. Spencer、Michael S.C. Thomas 和 James L. McClelland（编），

。