分析音频质量对使用自然情境长时录音开展婴儿导向语研究的影响

Mar´ıa Andrea CruzBland´on (maria.cruzblandon@tuni.fi)坦佩雷大学计算机科学系，邮政信箱 553，FI-33101，芬兰

Alejandrina Cristia (alecristia@gmail.com)D´认知研究系，巴黎高等师范学院、社会科学高等研究院、法国国家科学研究中心、巴黎文理研究大学，巴黎乌尔姆街 29 号，邮编 75005，法国

OkkoR¨as¨anen (okko.rasanen@tuni.fi)坦佩雷大学计算机科学系，邮政信箱 553，FI-33101，芬兰

arXiv:2305.01965v1 [cs.CL] 3 May 2023

摘要

早期语言习得建模旨在理解婴儿如何奠定语言能力的基础。建模涉及用于训练模型的输入数据的属性、待检验的认知假设及其算法实现，以及将模型与人类数据进行比较的评估方法。近期进展使计算模型能够使用更加自然情境的训练数据，也推动了对模型行为进行更自然情境化测试方法的开发。实现这一目标的关键一步，是建立具有代表性的语音数据集，其中包含婴儿在自然环境中听到的语音。然而，这类录音的一大缺点是通常带有噪声，目前尚不清楚声音质量会如何影响基于此类数据开展的分析和建模实验。本文探讨了这一问题，重点分析婴儿导向语（IDS）和成人导向语（ADS）。首先，我们对两个以儿童为中心的长时录音语料库（英语和法语）的语句进行了人工和自动音频质量标注。随后，我们比较了实验室数据集中以及自然情境数据的不同音频质量子集中的婴儿导向语和成人导向语的声学特征。最后，我们评估了音频质量和录音环境如何改变使用近期自监督学习模型进行建模分析所得的结论。结果表明，使用中等和高音频质量的自然情境语音数据，就声学分析和建模实验而言，关于婴儿导向语和成人导向语所得的结论大体相似。我们还发现，自动音质评估工具可用于筛选长时录音中适合进一步分析的片段，其结果与人工质量标注相当。

关键词：婴儿导向语；基准测试；长时录音；声学分析；计算建模；早期语言习得

引言

早期语言习得的计算建模旨在了解婴儿如何逐步建立语言技能。要开发可信的模型并将其用于认知假设检验，需要关注三个大体上相互独立的方面：模型本身、作为模型输入的语言数据，以及用于评估模型语言能力或表现的评估方案。

鉴于机器学习和数据收集方法的进步，计算模型已越来越能够从半自然的大规模听觉数据（Lavechin, de Seyssel, Titeux 等，2022年）或视听数据（Khorrami & R¨as¨anen，2021年）中学习。这与早期使用简化输入的研究形成对比，例如语音转写文本（Brent，1999年）、预先分段的语音单位（例如，de Boer & Kuhl，2003年），以及这些研究中常用的小规模语音语料库。这一进展很大程度上得益于所谓的自监督学习研究（SSL；例如，Oord、Li 和 Vinyals，2018年）。这些方法能够从低层次的感官输入中自主发现语音表征，而无需强烈的语言先验来引导学习过程。因此，它们也与语言学习的潜在模型密切相关。由此，近期研究已展示了仅凭声学语音对音位和词汇习得进行建模的成效（Lavechin, de Seyssel, Titeux 等，2022年），其中也包括从以儿童为中心的自然长时音频录音中学习的模型（Lavechin, de Seyssel, M´etais 等，2022年）。

在模型评估方面，目前也在持续推进工作，致力于制定更贴近现实的模型评估方案，以对应关于婴儿语言发展的实验或观察发现。例如，早期研究曾将模型与单项行为实验进行比较（Nixon & Tomaschek, 2021年），或分析发展模式（Lavechin, de Seyssel, Titeux, et al., 2022年；Nikolaus, Alishahi, & Chrupała, 2022年）。Cruz Blandon、Cristia 和 R¨as¨anen（2022年）最近也提出了一个评估框架，其中模型须接受与实验室中针对真实婴儿所进行的测试相当的计算测试，从而能够比较模型与婴儿的语言能力，并考察这种能力如何随模型训练数据和婴儿年龄而变化。

值得注意的是，行为研究及其计算复现通常会精心筛选刺激材料，以便研究人员控制研究所关注的刺激维度。但这可能会导致刺激材料不够真实、也比婴儿在现实生活中实际接触到的更简单。有证据表明，与更自然、自发的数据相比，使用实验室数据可能会高估语言能力的影响（MacDonald, R¨as¨anen, Casillas, & Warlaumont, 2020年）。因此，合乎逻辑的下一步是将计算模型评估扩展到更自然的数据，例如使用从儿童日常环境中的儿童中心长时录音中提取的音频刺激材料（例如，Bergelson 等，2019）。

然而，使用自然情境录音的一个核心问题是，与在受控环境中录制、质量更高的语料库相比，其音频质量变化很大，且平均而言非常低（例如，R{v1}as{v2}anen 等人报告的平均信噪比接近 0 dB，

2019年）。音频中的噪声可能影响声学分析和计算建模研究的结果，也可能导致用于训练和评估模型的数据集之间出现潜在的不匹配。因此，将从简化输入转向自然情境输入的影响，与从干净数据转向可能噪声更多的数据的影响区分开来十分重要。前者与推动使用更自然情境的数据有关，而后者在分析中只是一种潜在的干扰因素，除非所采用的声学分析方法和计算模型专门针对噪声进行了设计（这种情况并不常见）。

本文以婴儿导向语（IDS）和成人导向语（ADS）的分析为案例研究，旨在探讨上述问题。我们研究以下两个问题：1）自然语境下的 IDS 和 ADS 语音数据的基本属性是否不同于实验室中的刺激材料；2）基于自然语境数据的计算建模结果是否不同于基于实验室数据的结果。值得注意的是，我们在两项研究中均采用不同的声音质量筛选方法处理数据，以检验不同声音质量标准如何影响研究结果，并确定如何从长时录音中构建高质量刺激集的最佳实践。作为计算模型，我们采用了 Lavechin、de Seyssel、M´etais 等人（2022年）研究的一种用于统计学习的自监督算法。

方法论

本研究旨在构建一个包含婴儿导向语（IDS）和成人导向语（ADS）话语的自然情境测试集，用于评估语言习得计算模型在差异化处理 IDS 和 ADS 方面的能力，并研究该测试集的特性如何随不同声音质量标准的变化而变化，同时将其与实验室数据进行比较。

我们选择研究婴儿导向语（IDS），因为它在语言习得的起步过程中发挥着关键作用：它能引导婴儿将注意力集中在话语的相关部分，并使语音对比更容易辨识（Soderstrom，2007；Golinkoff、Can、Soderstrom 和 Hirsh-Pasek，2015）。研究发现，婴儿导向语广泛存在于不同语言和社群中，并具有一些共同特征；这些特征的显著程度因语言而异（Cox 等，2023）。与成人导向语（ADS）相比，婴儿导向语通常具有更显著的基频轮廓、更长的元音和更简单的语法（Soderstrom，2007；Hilton 等，2022）。此外，ManyBabies（MB）联盟提供了精心整理的实验室数据，其中包含婴儿导向语与成人导向语比较方面经过充分验证的研究发现（Frank 等，2020）；同时，也有按语音受话对象标注的长时录音（Bergelson 等，2019；Canault、Le Normand、Foudil、Loundon 和 Thai-Van，2016）。最后，早期有研究使用实验室和自然情境下的婴儿导向语数据进行建模，作者报告称，自然情境数据中的效应小于实验室数据中的效应（MacDonald 等，2020）。然而，作者在实验中并未控制声音质量，这可能解释了两类数据所得结果之间的差异。

长时段儿童中心录音通常比实验室录音噪声更大的原因在于，除了语音外，其中还包含背景噪声（如收音机或电视噪声）、重叠语音、婴儿哭声和尖叫声等其他噪声，而且录音距离从近场到远场不等。因此，无论对标注员还是计算模型而言，感知和处理长时录音都更具挑战性。因此，本研究在比较长时录音与实验室数据时，将音频质量作为可能影响结果的因素纳入考量。此外，我们还针对两种节律不同的语言开展分析：英语和法语。最后，我们也使用开发的测试集，借助自监督模型评估音频质量对模型评估的影响，以预测婴儿导向语和成人导向语话语（例如，参见 MacDonald 等，2020；R¨as¨anen、Kakouros 和 Soderstrom，2018 中较早的相关建模研究）。

用于基准测试长时段录音中婴儿导向语和成人导向语片段的语料库

为创建自然情境测试集中的婴儿导向语和成人导向语刺激材料，我们采用了 ACLEW 项目数据英语部分中经人工标注的婴儿导向语和成人导向语片段（Bergelson 等，2019年；Soderstrom 等，2021年）（下文简称 SEEDLings 数据），以及法语数据集中里昂项目数据经标注的片段（Canault 等，2016年）（下文简称里昂数据）。SEEDLings 数据和里昂数据均通过将 LENA 音频录音设备放置在婴儿的衣物中，连续录制一整天而采集。经标注的片段来自对婴儿完整录音中随机抽取的两分钟（SEEDLings）或十分钟（里昂）片段；SEEDLings 数据涉及十名婴儿（3 – 20 月龄），里昂数据涉及16名婴儿（3 – 41 月龄）。这些数据集的作者对录音片段中的话语[1] 及其说话者和受话者进行了人工标注。

在我们的实验中，我们将成年说话者（女性或男性）对成年听话者说出的语句归为成人导向语（ADS），将成年说话者对婴儿听话者说出的语句归为婴儿导向语（IDS）。据此，SEEDLings 数据共包含 1124 条成人导向语语句和 952 条婴儿导向语语句，里昂数据则包含 3915 条成人导向语语句和 863 条婴儿导向语语句。随后，我们按照下一小节中的说明，对这些语句进行了音质评估。

为了将长时录音中的话语与行为实验中使用的刺激材料进行比较，我们选择 MB 项目的刺激集作为参照，其中包含 120 条婴儿导向语话语和 120 条成人导向语话语（Frank 等，2020）。MB 录音使用 LENA 设备或移动设备录制，但均在固定的实验室环境中进行，录音对象为婴儿或成人，任务是描述物体（详情见 Frank 等，2020）。

[1]与采用基于 LENA 的自动处理、随后对标签进行人工核验的 MacDonald 等人（2020年）不同。

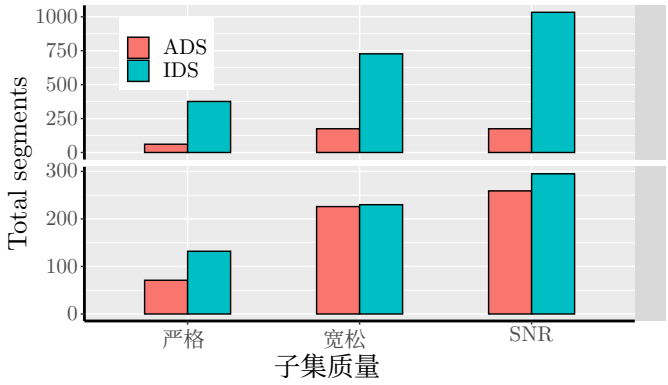


图1：里昂（上）和 SEEDLings（下）数据集中，婴儿导向语和成人导向语片段按照确定的片段质量分布。

评估婴儿导向语（IDS）和成人导向语（ADS）片段的音频质量

为确定自然情境下的婴儿导向语音和成人导向语音片段的音频质量，我们采用了两种标注策略：人工标注和自动标注。人工标注采用三个等级：质量好、中等和差。质量依据感知到的背景噪声、重叠语音、婴儿哭声和尖叫声，以及目标语音的响度和混响来界定。好与中等质量的区分依据是响度和与麦克风的距离（即，优质录音要求响度较高且混响较少）。一名标注员（第一作者）在数日内标注了两个数据集中的所有样本。

自动声音质量评分采用近期的信噪比（SNR）估算系统得出（Lavechin, M´etais, et al., 2022年）。首先估算所有话语的信噪比（单位为 dB），随后利用人工标注确定最适合将数据划分为良好质量和中等/较差质量数据的信噪比阈值。该流程得出的阈值分别为：SEEDLings 为 14.50 dB，里昂数据为 14.03 dB。

根据这些评分，我们为实验分析定义了三个数据子集：严格质量，由人工标注为良好质量的话语组成。宽松质量，由人工标注为良好和中等质量的话语组成；以及信噪比质量，其自动估计值高于设定的信噪比阈值的话语。图 1 显示了不同质量子集中婴儿导向语和成人导向语话语的样本数量。从图中可以看出，声音质量与测试集大小之间存在权衡，信噪比方法的质量要求最宽松，样本数量也最多。

婴儿导向语（IDS）与成人导向语（ADS）特征的声学分析

根据 MacDonald 等人（2020年），我们进行了声学分析，以比较长时录音和实验室录音在基本婴儿导向语/成人导向语特征方面的差异。为此，提取基频对数的均值（对数基频）和对数基频变异性（对数基频标准差），用于所有话语。使用 PRAAT[2]进行语音分析。此外，面向婴儿的语音往往比面向其他成人的语音更短（Soderstrom，2007）。因此，我们在分析中纳入了话语时长（以秒为单位）。最后，有证据表明，说话者会对婴儿导向语采用夸张的发音（Cox 等，2023），以帮助婴儿处理信息。尽管无法从未经语音标注的数据中提取元音空间测量值，但我们使用平均频谱倾斜度 作为一般发声力度的替代指标，该指标也能反映说话风格的潜在差异。为此，我们使用了第一个梅尔频率倒谱系数（MFCC）。

总体而言，我们预期两种语言的结果都能复现 MacDonald 等人（2020年）的发现，即婴儿导向语话语的基频和变异性高于成人导向语话语。对于时长，我们预期实验室录音中的时长更长，且在自然情境数据中不受子集质量影响。对于频谱倾斜度，我们预期发声力度较大的话语具有更大的频谱倾斜度。我们使用独立样本 t 检验来判断婴儿导向语和成人导向语的声学特征是否存在显著差异，并报告了差异的大小，采用 Cohen 的*d*（Lipsey 和 Wilson，2001年）衡量。

计算建模分析

建模实验旨在检验：1）基于自然情境数据开展的实验，是否会复现基于实验室数据观察到的行为；2）数据的质量筛选是否会影响这一结果。虽然未必能预期前者成立（这正是最初使用自然情境数据的原因），但了解测试数据的声音质量是否会显著影响模型行为非常重要。

用于语音表征学习的自监督模型我们使用了（Lavechin、de Seyssel、M´etais 等，2022年）提出的自监督统计学习神经网络模型，将其作为早期语言习得的模型。该模型以无监督的方式直接从原始语音信号中学习。学习过程的核心是优化一个预测过程，该过程尝试根据当前语音输入预测未来的语音观测。然而，模型并不直接预测声学信号，而是学习预测未来语音自身的潜在表征，使其也能够根据预测任务的需要自行构造这些表征（参见大脑根据感觉输入预测自身的神经激活，而非实际的物理信号）。建议读者参阅 Oord 等人（2018年），了解学习机制的技术细节。

模型使用来自 LibriVox 和 LibriLight 的朗读有声书进行训练，并分别针对英语和法语训练模型（Kearns, 2014；Kahn 等，2020）。我们使用了 Lavechin、de Seyssel、M´etais 等人（2022）的模型版本，训练时长分别为50、100、200、400、800、1600或3200小时，以模拟学习者不同程度的语言经验。

[2]通过 Python 库封装：Parselmouth（Jadoul、Thompson 和 de Boer，2018年）

可预测性分析 语音的可预测性与语音结构的编码有关。能够更准确预测输入的模型，对语音信号的理解也更好；而难以预测输入的模型，则可以被视为对刺激材料更感到“意外”。先前研究发现，与 ADS 相比，IDS 的可预测性较低，这与对 IDS 更强的注意力偏好有关（R¨as¨anen 等，2018年；MacDonald 等，2020年；另见 Cruz Blandon 等，2022年）。然而，本文不对注意力中的熟悉性／新颖性编码采取立场，而只是将输入的可预测性用作衡量模型对给定语音风格的掌握程度的指标。这样，我们便可以比较不同数据集在自然度和噪声水平方面可能存在的不匹配。

对于这些在有声读物上训练的模型而言，它们并未接触过婴儿导向语或成人导向语言语风格的示例。然而，基于先前的研究结果（Lavechin、de Seyssel、M´etais 等，2022年），我们预计，随着训练数据量的增加，模型将获得更好的语音表征和更高的预测准确率。如果模型学习到的语音表征涵盖了不同言语风格中普遍存在的特征，我们预计婴儿导向语和成人导向语都具有相似的可预测性。另一种可能是，成人导向语话语的可预测性高于婴儿导向语，因为成人导向语与模型在训练期间接触到的朗读语音更为相似。如果测试子集的质量不影响模型行为，我们预计三种质量条件下都会呈现相同的可预测性模式。我们预计，实验室内数据与长时段数据之间的可预测性结果可能有所不同。这是因为模型关注的是完整的声学信号，而不是一小组经过选择的特征，因此，从语音感知的角度来看，声学测量结果相似的数据仍可能存在显著的结构差异。

结果

声学分析

表 1 展示了两种言语风格、两个语料库、不同声音质量子集以及作为参照的 MB 数据所获得的声学特征测量结果。对于三种声学特征，MB 数据符合文献中的预测（另见 MacDonald 等，2020）：与成人导向语相比，婴儿导向语的 F0 更高、时长更短、谱倾斜更高。

至于实验室数据与自然情境数据，MB数据中声学特征差异的效应量（Cohen’s d）总体上大于长时录音，这也重现了 MacDonald 等人（2020年）的发现。这进一步证明，与家庭录音相比，实验室录音中 IDS 与 ADS 的差异往往更为显著。然而，观察 SEEDLings 和里昂数据中呈现的模式时，只有 SEEDLings 数据集重现了 MB数据中的大多数发现；谱倾斜除外，因为 IDS 与 ADS 之间未发现显著差异；时长方面也并非如此，只有宽松质量重现了这一模式，了这一模式。与英语数据不同，法语数据（里昂）中 ADS 语句的平均 F0 和频谱倾斜度更高，而 IDS 语句比 ADS 语句更长。

里昂语料库的结果与此前的发现形成对比，此前的研究发现，婴儿导向语的基频高于成人导向语（Cox 等，2023年）。在人工检查音频质量时，我们发现，有些话语被标注为成人导向语，但实际上成人是在与房间里的另一个孩子交谈。标注中的这种不一致可能会影响测量结果的整体准确性。然而，这无法解释婴儿导向语和成人导向语中声学线索值为何出现反转。因此，目前尚不清楚这一结果的原因；这也说明，对自然语料的分析可能得出与基于控制程度更高的数据集进行分析时预期不同的发现。

关于各子集的音频质量，我们观察到，在两个自然语料库中，同一声学特征的测量值在三种音频质量之间略有差异。总体而言，数据越干净，效应量越大（见表1中的 Cohen’s d）。不过，在严格音频质量下观察到的主要效应及其方向，很可能也会出现在质量筛选要求较低的情况下。因此，从声学分析的角度来看，录音质量并不会显著改变所研究特征的性质。由于手动和自动质量筛选所得结果之间没有明显差异，这些发现还表明，自动筛选适用于长时音频数据的基础声学分析。我们将进一步探究这一发现是否同样适用于使用计算模型进行的可预测性分析。

建模分析

图2展示了在<样式 id='1'>严格质量</样式>下，SEEDLings、里昂和 MB数据的自监督模型的预测准确率。结果分别针对7种训练数据量（检查点）展示。图3和图4展示了 SEEDLings 和里昂数据在<样式 id='3'>宽松</样式>和<样式 id='5'>信噪比</样式>质量下的结果。

实验室MB数据的结果表明，模型在仅经过50小时训练后，准确率就已达到高于随机水平（> 0.02）；随着模型使用更多训练数据，其可预测性也随之提高；无论训练所使用的语言是哪一种，结果均如此。此外，与婴儿导向语（IDS）话语相比，模型对成人导向语（ADS）话语表现出更高的可预测性。然而，随着训练量增加，实验室数据中这两种语体之间的差异会逐渐缩小。

将 MB数据与严格质量子集中长时录音的数据进行比较（图 2），我们发现，SEEDLings 数据和里昂数据仅复现了准确率随数据增加而上升的趋势。总体而言，训练数据越多，可预测性越高，这同样不受训练所用语言的影响。然而，对于 SEEDLings 数据，婴儿导向语与成人导向语之间没有显著差异。对于里昂数据，趋势与 MB数据相反：与成人导向语话语相比，模型对婴儿导向语话语的可预测性更高。这些结果表明，基于实验室数据进行建模的结果

表 1：声学分析结果：报告了三个音频质量子集中婴儿导向语和成人导向语片段的平均音高（F0）、以对数尺度表示的基频变异性、谱倾斜和时长（秒）。效应量以 Cohen's d 表示，星号表示所有显著差异（在 $\alpha = 0.05$）。在婴儿导向语和成人导向语之间存在显著差异时，较大的数值以粗体标示。

		平均对数基频			对数基频标准差			谱倾斜			时长（秒）		
		IDS	ADS	d	IDS	ADS	d	IDS	ADS	d	IDS	ADS	d
SEEDLings	严格	**5.31**	5.07	0.80*	0.22	0.20	0.16	135.71	137.23	0.04	1.22	1.46	0.27
	宽松	**5.32**	5.10	0.77*	0.23	0.22	0.07	133.67	133.75	0	1.37	**1.86**	0.43*
	SNR	**5.39**	5.16	0.81*	0.23	0.23	0.03	135.72	135.09	0.02	1.12	1.20	0.10
Lyon	严格	5.46	**5.58**	0.43*	0.25	0.22	0.20	82.42	85.09	0.07	2.25	1.99	0.21
	宽松	5.45	**5.50**	0.18*	0.25	0.26	0.05	81.98	**90.81**	0.23*	**2.51**	2.24	0.19*
	SNR	5.46	**5.56**	0.35*	0.25	0.24	0.06	84.91	**91.52**	0.17*	**2.15**	1.86	0.24*
ManyBabies		**5.39**	5.19	1.18*	0.28	0.29	0.03	**180.16**	169.08	0.31*	2.97	**5.14**	1.00*

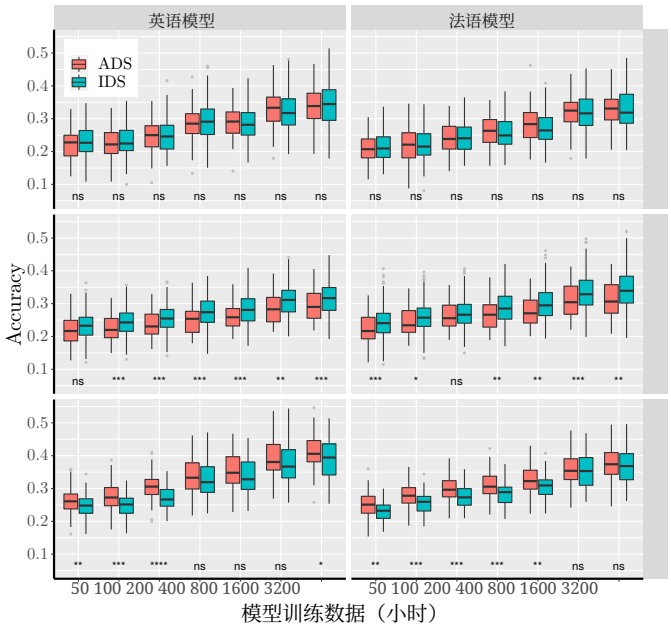


图2：不同模型训练数据规模（x轴）下严格音频质量数据上的预测准确率。顶部：SEEDLings 数据。中部：里昂数据；底部：MB数据。左侧：英语训练模型；右侧：法语训练模型。红色为ADS，绿色为ADS。IDS与ADS之间的显著差异以星号表示，无差异则以ns表示（$*$代表$p < 0.05$，$**$代表$p < 0.01$，$***$代表$p < 0.001$）。

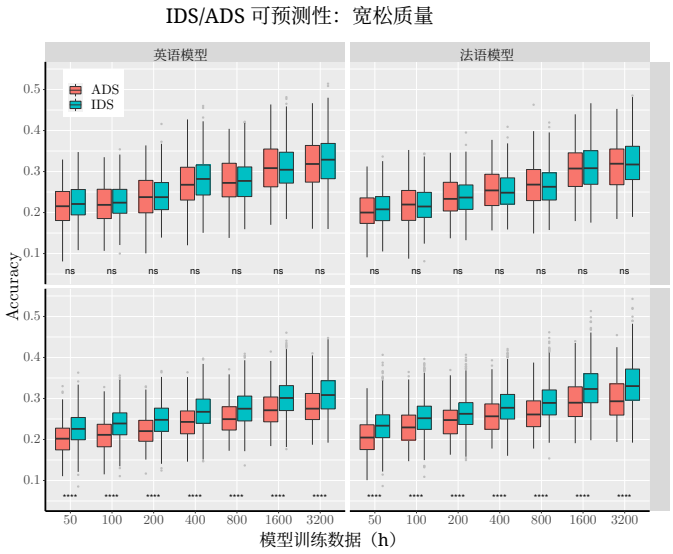


图3：宽松质量子集中的 IDS 和 ADS 语句的可预测性。有关图表详情，请参见图2的图注。

与长时数据的结果可能不同，我们将在最后一节进一步讨论这一点。

在分析三种音频质量下的结果时（见图3和图4），我们可以看到，各语料库中发现的可预测性模式在三种质量条件下都适用。与声学分析一样，随着噪声数据增多，可预测性会略微下降（例如，对于 SEEDLings 数据的严格子集，50小时英语模型的准确率从 ≈ 0.24 降至宽松子集的 ≈ 0.22，再降至信噪比子集的 ≈ 0.20）。然而，婴儿导向语和成人导向语的相对可预测性随训练数据量变化的基本模式不受声音质量影响。

总体而言，尽管该模型是在有声读物上训练的，

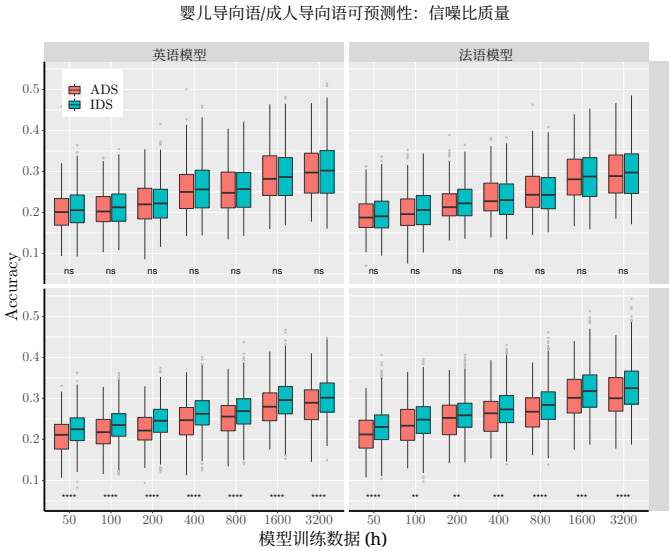


图4：信噪比质量子集中的婴儿导向语和成人导向语话语的可预测性。有关图表的详细信息，请参见图2的图注。

它仍能够以相近的准确率预测两种言语风格，在预测任一种风格时都没有显著差异（IDS 与 ADS 之间的差异不超过 0.1）。结果表明，该模型学习到了适应口语特性的语音表征，但似乎并未以特别不同的方式处理 IDS/ADS 风格。无论模型是在英语还是法语上训练，这一现象都同样存在，不过对非母语语音的预测准确率略低。这可能是因为模型并非在自然的 ADS 或 IDS 上训练，而是在朗读有声书上训练；因此，从朗读语音转向自然语音时，语音特性可能出现程度相近的不匹配，无论自然语音是 IDS 还是 ADS。

讨论与结论

为了构建更贴近自然情境的计算建模评估环境，有必要了解实验室录音与家庭录音之间的差异，以及不同数据集中的声音质量差异所带来的影响。本文以 IDS/ADS 对比为例，提取了长时录音的三种音频质量指标，并将其声学特征与经过筛选、用于实验室测试的录音特征进行对比。我们还使用可预测性分析，探究音频质量是否会影响建模结果。

总体而言，实验室数据与长时录音数据之间存在若干差异：录音的内容和时长（短时、受控的录音环节与不受控制的全天日常活动）、话语所处的社会语境（另见 MacDonald 等，2020），以及实验室数据中婴儿导向语与成人导向语之间的声学差异更大，而长时录音数据中的差异则较小。这表明婴儿在自然情境中面对的输入，可能包含较之实验室情境中所遇到的特征更不显著的特征。因此，无论是声学分析还是计算建模结果，实验室数据中发现的言语风格差异都未必能在家庭数据中重现（例如，婴儿导向语与成人导向语的谱倾斜没有差异）。即使声学模式相似（例如，MB 和 SEEDLings 数据中的基频及其变异性），实验室数据和家庭数据在评估计算模型时也可能得出不同的结论（例如，参见可预测性分析）。

然而，不同说话风格的显著性在自然情境中也可能存在更大差异，甚至很可能如此。自然情境中 IDS 的平均声学属性，未必能反映这些交际场景中儿童接收到的语言输入，甚至未必能反映话语中那些真正需要通过风格手段控制注意力、促进学习、注意力控制或情感调节的特定时点（另见 MacDonald 等，2020 年的讨论）。在为自然情境下的模型创建评估测试集时，解释结果需要考虑上述方面。归根结底，测试方案的目的也很重要：自然情境数据可用于检验模型在“真实环境”中的行为，并有助于提出关于现实学习场景中 IDS 和 ADS 的作用及特征的新实证假设。然而，在此之前，应先基于存在可比较婴儿行为数据的数据进行 模型验证；目前，这类数据主要局限于受控实验中的实验室数据。

关于音频质量，我们的结果表明，通过对片段进行人工标注，或使用自动化工具进行信噪比估计，都可以获得相当且可接受的音频质量（至少对于本文考察的声学特征和建模任务而言）。实际意义在于，自动化工具（例如 Lavechin、M´etais 等人（2022年）开发的工具）可以有效地筛选出符合所需音频质量要求的数据子集。当与其他用于处理长时段数据集的自动化工具结合使用时，这一点尤其有用，因为人工质量标注无法扩展到数百或数千小时的语音数据。

总而言之，本研究增进了我们对录音条件（实验室内与家庭环境）如何影响分析和计算建模结果的共同认识；本研究考察的对象是婴儿导向语（IDS）和成人导向语（ADS）语音。理想情况下，同时使用实验室内数据和家庭数据，可以更全面地分析所关注的现象。此外，借助自动化工具即可确定长时录音的音频质量，无需进行繁琐的标注，而且不太可能遗漏在人类听者标注为高质量的数据中同样存在的重要发现。未来的研究需要更全面地检验模型训练数据与测试数据之间潜在的领域不匹配，例如，也可以使用自然语料而非干净数据来训练模型。不过，这将留待未来研究。

致谢

作者 MACB 和 OR 获得了芬兰科学院拨款（编号：314602、320053 和 345365）以及 Kone基金会 L-SCALE 项目的资助。AC 获得了法国国家研究署（ANR-16-DATA-0004 ACLEW）、欧洲研究委员会（ExELang，101001095）以及 J. S. McDonnell 基金会「理解人类认知学者奖」的资助。我们感谢 Marvin Lavechin 和 Maureen de Seyssel 在研究早期提出的见解，这些见解对本研究有所裨益。测量数据和脚本可在 https://github.com/SPEECHCOG/ids 获取。音频质量分析

参考文献

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A.（2019年）。北美婴儿听到了什么？一项大规模跨语料库分析。《发展科学》，*22*(1), e12724。doi: https://doi.org/10.1111/desc.12724 Brent, M. R.（1999年）。一种用于切分和词语发现的高效且概率可靠的算法。《机器学习》,*34*, 71–105。doi: 10.1023/A:1007541817488 Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H.（2016）。欧洲法语语言环境分析系统（lenaTM）的可靠性。《行为研究方法》, *48*, 1109-1124。doi: 10.3758/s13428-015-0634-8 Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roep- storff, A., Bryant, G., & Fusaroli, R.（2023年）。婴儿导向语声学特征的系统性综述与贝叶斯荟萃分析。*Nature Human Behaviour*, *7*, 114–133。doi: 10.1038/s41562-022-01452-1 Cruz Blandon, M. A., Cristia, A., & R¨as¨anen, O.（2022年）。基于荟萃分析中稳健实证数据评估婴儿语言发展计算模型：内容、原因与方法？*PsyArXiv*。（版本1）doi: 10.31234/osf.io/yjz5a de Boer, B., & Kuhl, P. K.（2003年）。利用计算机模型研究婴儿导向语的作用。《在线声学研究快报》, *4*, 129–134。doi: 10.1121/1.1613311 Frank, M. C., Alcock, K. J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S., . . . Soderstrom, M.（2020年）。利用婴儿指向语偏好量化婴儿期研究中的变异性来源。《心理科学方法与实践进展》, *3*(1), 24– 52。doi: 10.1177/2515245919900809 Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh- Pasek, K.（2015年）。（宝贝）跟我说话：婴儿导向语的社会语境及其对早期语言习得的影响。《心理科学当前方向》, *24*(5), 339-344。doi: 10.1177/0963721415595345 Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., . . . Mehr, S. A.（2022年）。不同……中的婴儿导向语和歌曲的声学规律性文化。*NatureHuman Behaviour*，2397-3374。doi: 10.1038/s41562-022-01410-x Jadoul, Y., Thompson, B., & de Boer, B. (2018). 介绍 Parselmouth：一个用于 Praat 的 Python 接口。*Journalof Phonetics*，*71*，1–15。doi: https://doi.org/10.1016/j.wocn.2018.07.001 Kahn, J., Rivi`ere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazar´e, P., . . . Dupoux, E. (2020). Libri-light：一项针对有限监督或无监督自动语音识别的基准测试。载于 *Icassp2020 - 2020 ieee international conference on acoustics, speech andsignalprocessing (icassp)*（第 7669-7673 页）。doi: 10.1109/ICASSP40776.2020.9052942 Kearns, J. (2014). Librivox：免费的公有领域有声读物。载于*Emerald group publishing limited.*Khorrami, K., & R¨as¨anen, O. (2021). 音素、音节和词能否作为跨情境视听学习的副产品涌现？——一项计算研究。*Language DevelopmentResearch*，*1*，123-191。doi: 10.34842/w3vw-s845 Lavechin, M., de Seyssel, M., M´etais, M., Metze, F., Mohamed, A., Bredin, H., . . . Cristia, A. (2022). 早期语音习得的统计学习模型难以处理以儿童为中心的音频数据。*PsyArXiv*。（版本 4）doi: 10.31234/osf.io/5tmgy Lavechin, M., de Seyssel, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2022). 统计学习为早期语言习得奠定基础。*PsyArXiv*。（版本 2）doi: 10.31234/osf.io/rx94d Lavechin, M., M´etais, M., Titeux, H., Boissonnet, A., Copet, J., Rivi`ere, M., . . . Bredin, H. (2022). Brouhaha：针对语音活动检测、语音噪声比和 C50房间声学估计的多任务训练。doi: 10.48550/ARXIV.2210.13248 Lipsey, M. W., & Wilson, D. B. (2001). 实用元分析。美国加利福尼亚州千橡市：Sage Publications, Inc. MacDonald, K., R¨as¨anen, O., Casillas, M., & Warlaumont, A. S. (2020). 测量儿童家庭语言环境中的韵律可预测性。载于 第*42*届认知科学学会年会论文集 （第 695-701 页）。认知科学学会。Nikolaus, M., Alishahi, A., & Chrupała, G. (2022). 与小猪佩奇一起学习英语。计算语言学协会汇刊，*10*，922–936。doi: 10.1162/tacl a 00498 Nixon, J. S., & Tomaschek, F. (2021). 早期婴儿语音学习中的预测与误差：一种语音习得模型。*Cognition*，*212*，104697。doi: 10.1016/j.cognition.2021.104697 Oord, A. v. d., Li, Y., & Vinyals, O. (2018). 基于对比预测编码的表征学习。计算研究资料库，*abs/1807.0*。R¨as¨anen, O., Kakouros, S., & Soderstrom, M. (2018). 婴儿导向语之所以有趣，是因为它令人惊讶吗？——将婴儿导向语的特性与韵律层面的统计学习和注意力联系起来。*Cognition*，*178*，193-206。doi:

https://doi.org/10.1016/j.cognition.2018.05.015 R¨as¨anen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., . . . Soderstrom, M. (2019). 使用语音的语言无关音节划分方法，根据不同语言环境中的儿童全天候中心录音自动估算词数。*Speech Communication*，*113*，63-80。doi: https://doi.org/10.1016/j.specom.2019.08.005

Soderstrom, M. (2007). 超越儿语：重新评估婴儿尚未开始说话前所接收语音输入的性质与内容。*Developmental Review*，*27*(4)，501-532。doi: 10.1016/j.dr.2007.06.002

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). 开发用于研究婴儿真实世界语言经验的跨文化标注系统与元语料库。*Collabra: Psychology*，*7*(1)，23445。