arXiv:2301.07829v2 [cs.HC] 2 Aug 2023

警告：人类无法可靠地检测语音深度伪造

Kimberly T. Mai[1,2,*], Sergi Bray[1,2], Toby Davies[1], Lewis D. Griffin[2]

1 伦敦大学学院安全与犯罪科学系，英国伦敦

2 伦敦大学学院计算机科学系，英国伦敦

* 通讯作者

电子邮箱：kimberly.mai@ucl.ac.uk (KTM)

摘要

语音深度伪造是由机器学习模型生成的人工语音。以往文献指出，由于人工智能的发展可能被滥用，深度伪造已成为由此产生的最大安全威胁之一。然而，研究人类检测能力的研究仍然有限。我们向 $n = 529$ 名个人播放了真实音频和深度伪造音频，并请他们辨别其中的伪造内容。我们分别用英语和普通话进行了实验，以了解语言是否会影响检测性能和决策依据。我们发现，检测能力并不可靠。听众仅在 73% 的情况下正确识别出深度伪造，而且两种语言的可检测性没有差异。通过提供语音深度伪造示例来提高听众的认知度，只能略微改善检测结果。随着语音合成算法不断进步，生成的语音也越来越逼真，我们可以预见，检测任务将变得更加困难。检测语音深度伪造的难度证实了其被滥用的可能性，也表明我们需要采取措施来防范这一威胁。

引言

对手已经在利用语音深度伪造实施诈骗。2020年，香港一位银行经理接到一通电话，对方听起来像是他此前交谈过的一位公司董事 [1]。这名所谓的董事要求银行经理批准总额达 3500 万美元的转账。基于双方已有的关系，银行经理转出了 40 万美元，直到后来才意识到情况不对。这位银行经理成了一场精心策划的骗局的受害者：诈骗分子利用深度伪造技术克隆了该董事的声音。这并非孤立事件。2019年，一家英国公司的首席执行官遭到其经理的语音深度伪造欺骗，向一家匈牙利供应商转账 22 万欧元 [2]。

语音深度伪造是由机器学习模型生成的人工语音。随着研究迅速进展，如今只需少量音频样本 [3]，就能生成听起来逼真的克隆语音。这一进展意味着语音深度伪造可能被用于各种犯罪活动。除了冒充之外，犯罪分子还可能利用深度伪造进行鱼叉式网络钓鱼、传播假新闻，以及绕过生物识别认证系统 [4–6]。

现有的语音深度伪造检测研究侧重于在语音认证情境下开发机器学习系统 [7–9]。超出这一生物识别应用场景的比较，以及衡量人类检测能力的研究都十分有限 [10]。

现有研究的现状引发了一些问题。首先，机器学习系统需要大量数据进行训练 [11]，而且难以解释 [12]。在分析这些系统时，我们不清楚哪些特征能够区分合成语音与真实语音。因此，了解人类用什么来识别深度伪造，有助于更好地理解黑箱机器学习系统的工作原理。

其次，聚焦于自动化生物识别认证，并不能量化语音深度伪造在其他潜在犯罪应用中的威胁。了解这些威胁的程度至关重要。多项研究认为，语音深度伪造的其他用途更令人担忧，例如通过声音冒充误导他人 [5, 6]。专家预计，深度伪造产生的虚假信息会在多个层面侵蚀信任：针对个人、组织，乃至整个社会 [13]。此外，据估计，到 2026 [14], ，多达 90% 的在线内容将由合成方式生成，这意味着

内容审核将面临挑战。因此，了解语音深度伪造的风险，有助于制定更有效的防御措施和法规，在危害发生前加以遏制。

我们通过测量人类区分真实语音与合成语音的能力，试图回答这两个问题。我们开展了一项在线实验，让个人听取真实语音和伪造音频片段，并尝试分辨二者。

我们将参与者随机分配到两种实验设置中。在第一种设置中，我们每次向参与者播放一个音频片段，并请他们判断该片段是否为伪造内容。在第二种设置中，我们向参与者播放包含相同语音的音频片段对（一段是真实语音，另一段是合成语音），并请他们找出合成音频。

我们使用英语和普通话开展实验，以了解听众是否利用语言特定属性检测深度伪造，并观察深度伪造检测在某种语言中是否比在另一种语言中更容易。最后，我们加入了随机干预，以评估让参与者熟悉语音深度伪造示例是否能提升检测性能。

结果表明，听众的检测能力有限，而且他们在不同语言中的表现相近。此外，让参与者熟悉相关示例虽能提升表现，但提升幅度很小。

背景

深度伪造媒体

深度伪造是以某个人的形象制作的合成媒体，属于生成式人工智能（人工智能）领域。生成式人工智能是机器学习算法的一个子集，这些算法会学习数据集 [11] 的模式和特征。算法利用这些知识生成与原始数据相似的合成内容。深度伪造特指生成式人工智能生成的、与人类及其行为相似的内容。

深度伪造媒体有不同的模态：

1. **图像**：这种模态包含使用各种技术生成的静态人脸。这些技术包括：

    - **从零开始生成**：生成对抗网络 [15] 或扩散模型 [16] 合成出一个虚构身份。
    - **人脸融合**：将长相相似的人脸融合起来，生成包含源人脸特征的身份 [17]。
    - **人脸替换**：将源人脸替换到另一张图像中的目标人脸位置 [18]。

2. **视频**：这种模态呈现个人执行动作的画面。目前，用于合成视频的技术与图像合成技术相似。图像合成技术应用于每一帧，再将各帧拼接起来形成视频。

3. **语音**：这种模态传递的信息听起来像是真人的声音。尽管音频可以指一般的声音合成，但在学术文献中，“音频”“语音”和“声音”深度伪造这几个术语可以互换使用。为确保全文术语一致，我们将其称为“语音深度伪造”。

此外，深度伪造要么模仿已知身份（**定向**），要么与熟悉的身份并不相似（**非定向**）。例如，我们可以将针对政治人物制作的视频深度伪造归类为定向。相反，从零开始创建、且未经过条件设定来模仿某个特定个人的通用人脸，则属于非定向。

有关深度伪造术语的更多信息，请参阅 Zhang（2022） [19] 。由于关注语音深度伪造的研究较少，我们将重点讨论这种模态。

合成语音

生成模型常用于合成语音。使用生成模型的语音合成器遵循一个通用框架：

1. **数据收集**：收集说话人的若干音频录音。

2. **预处理**：将音频录音转换为替代格式，以便生成模型更轻松地处理这些录音。

3. **训练**：将处理后的音频录音输入生成模型，使其学习数据的模式和特征。训练完成的模型通常称为声码器。

这些框架通常包含文本转语音（TTS）模块，以便更轻松地生成语音。在这种设置下，生成模型还会看到与音频录音相对应的文本转录。

我们在图1中展示了该框架的可视化图示。

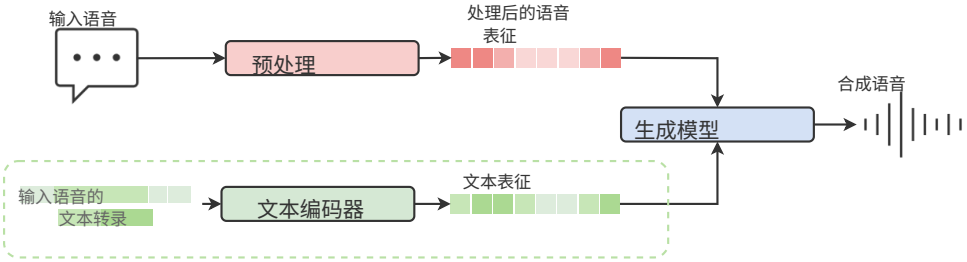

图1. 典型生成式语音合成模型示意图。

人类深度伪造检测能力的相关研究

大多数深度伪造检测研究考察人类表现时，使用的都是视觉媒体。面对涉及政治人物的深度伪造内容，参与者会依靠政治素养形式的语境知识来识别伪造内容 [20, 21]。

如果去除这类背景知识，检测任务就会更加困难。多项图像研究表明，人类的表现并不比随机猜测好多少 [22, 23]。评估视频时也没有任何改善 [24–26]。此外，这些研究表明，人类对自己的深度伪造检测能力过于自信 [25]。

上述多项研究探讨了干预措施能否提升检测性能。然而，这些干预措施的效果仍有争议。Bray 等人 [22] 通过在主要任务开始前展示深度伪造示例，让参与者熟悉深度伪造。作者还提醒参与者注意伪造图像中常见的错误。尽管这些干预措施提升了深度伪造检测性能，但也加剧了整体的怀疑态度，因为更多真实图像被误判。还可以指出，指出这些错误会使参与者产生偏见，且

妨碍他们独立识别深度伪造的典型特征。K¨obis 等人 [25] 通过告知参与者深度伪造的影响并奖励猜测正确者来实施干预措施。两种干预措施都未能提升表现。

相比之下，其他作者发现，根据机器学习模型输出制定的干预措施能够提升检测能力。Tahir 等人 [26] 借助机器学习可解释性工具，制作了介绍伪造图像特征的教育材料。作者发现，与最初的对照组相比，检测性能有所提升。然而，近期一项研究 [27] 对这些工具的可靠性提出质疑，因为作者证明，输出可视化结果可能受到操纵。Groh 等人 [24]允许参与者在查看机器学习模型的预测结果后修改自己的选择。这种合作方式显著提升了结果。

研究人类检测语音深度伪造能力的工作较少。Watson 等人 [28] 向大学生播放了八段片段，并请他们判断这些片段是真实还是伪造的。他们发现，较短的片段更容易识别。不过，该研究的样本量较小，且参与者主要是年轻、受过大学教育的人群。ASVspoof 挑战赛的组织者开展了一项样本量更大的实验 [29]。他们请 1,145 名参与者设想自己在呼叫中心工作，并判断打入的电话是由人类还是人工智能发出的。不过，该实验仅限于说话人验证场景。

M¨uller 等人 [30] 举办了一场游戏，378 名参与者与一个机器学习模型竞赛，判断音频片段是否为伪造。与 Groh 等人 [24], 的研究类似，他们发现，机器学习模型提供的反馈提升了人类表现。在他们的实验中，M¨uller 等人 [30] 发现，人类与人工智能准确率之间的差距约为 10%。然而，他们的研究仅使用英语片段，每次只向参与者播放一个音频片段，也没有收集参与者信心程度的信息。

我们在表1中总结了相关文献。我们注意到，Barari 等人 [20]在分析中提到了伪造语音刺激材料。然而，他们使用演员来制作语音，而非生成式人工智能。因此，我们未将其纳入分析。

表1. 测量人类检测深度伪造能力的相关文献摘要。

模态	Year	作者	深度伪造刺激材料
图像	2021	Nightingale & Farid [23]	使用 StyleGAN 生成的人脸2 [31]
	2022	Bray 等人 [22]	使用 StyleGAN2 生成的人脸
视频	2021	Barari 等人 [20]	政治人物的换脸视频
	2021	Groh 等人 [24]	来自深度伪造检测挑战数据集的换脸视频 [32]
	2021	K¨obis 等人 [25]	来自深度伪造检测挑战数据集的换脸视频
	2021	Tahir 等人 [26]	来自 Celeb-DF 的换脸视频 [33], FaceForensics++ [34] 和 DeepFaceLab [35]
	2022	Appel 与 Prietzel [21]	政治人物的换脸视频
语音	2020	Wang 等人 [29]	在 ASVspoof2019 中使用文本转语音和语音转换系统生成的欺骗性话语
	2021	Watson 等人 [28]	使用 MelGAN 生成的音频片段 [36]
	2022	M¨uller 等人 [30]	来自 ASVspoof2019 数据集的欺骗性话语 [29]

材料与方法

我们的研究问题如下：

1. 人类检测语音深度伪造的能力如何？

2. 检测能力是否因语言不同而有所差异？

3. 以示例和补充背景信息为形式的干预措施，能否提高检测性能？

我们的实验关注的是人类表现，而非自动检测器的表现。通过这种实验设置，我们能够量化人类接触语音深度伪造时所面临的威胁。

刺激材料

真实刺激材料

我们从两个公开可用数据集中收集了真实刺激材料。这两个数据集均包含一位女性说话人朗读通用句子的录音。数据集还提供了音频的文本转录。我们选择这类数据集，以防止参与者在检测任务中利用外部线索。

我们使用 LJSpeech [37] 作为英语数据集。该数据集包含一位说话人朗读七本非虚构书籍中的段落，时长从一秒到十秒不等。

我们使用中文标准普通话语料库（CSMSC） [38] 作为普通话数据集。该数据集中使用的语料旨在尽可能全面地涵盖普通话声调和韵律。

深度伪造刺激材料

为了生成深度伪造刺激材料，我们使用了在两个数据集 [39] 上训练的公开可用文本转语音模型。具体而言，我们选择了预训练的 VITS 模型 [40]。VITS 是一种端到端文本转语音模型，它将数据预处理和声码器整合到一个统一框架中。

我们从两个数据集的验证集划分中随机选取了 50 个句子来生成深度伪造内容。真实刺激材料也使用相同的句子。因此，我们总共得到了 100 个片段。

验证集划分由未用于训练机器学习模型的样本组成。使用未见数据是良好的做法，因为这能反映训练后的模型具有多强的泛化能力。因此，生成音频应包含我们预期会从机器学习模型中听到的伪影。这些伪影可能成为区分深度伪造的有效特征。如果使用训练期间已见过的样本，模型可能会完美模仿这些样本，从而无法产生具有代表性的伪影。

实验流程

英语和普通话实验的设置完全相同。我们将参与者随机分配到两种配置：一元和二元。在两种配置中，我们都要求参与者用十分制李克特量表评估自己选择的信心程度，并提供自由文本理由。参与者可以随意反复收听这些片段。我们没有向参与者提供反馈，告知他们的选择是否正确。与 M¨uller 等人和 Groh 等人 [24, 30], 所描述的设置相比，不提供反馈能营造出更真实的情境。在现实生活中遇到语音深度伪造时（例如，通过诈骗电话），人们并不知道这些声音是伪造的。我们在图2中展示了两种配置的截图。

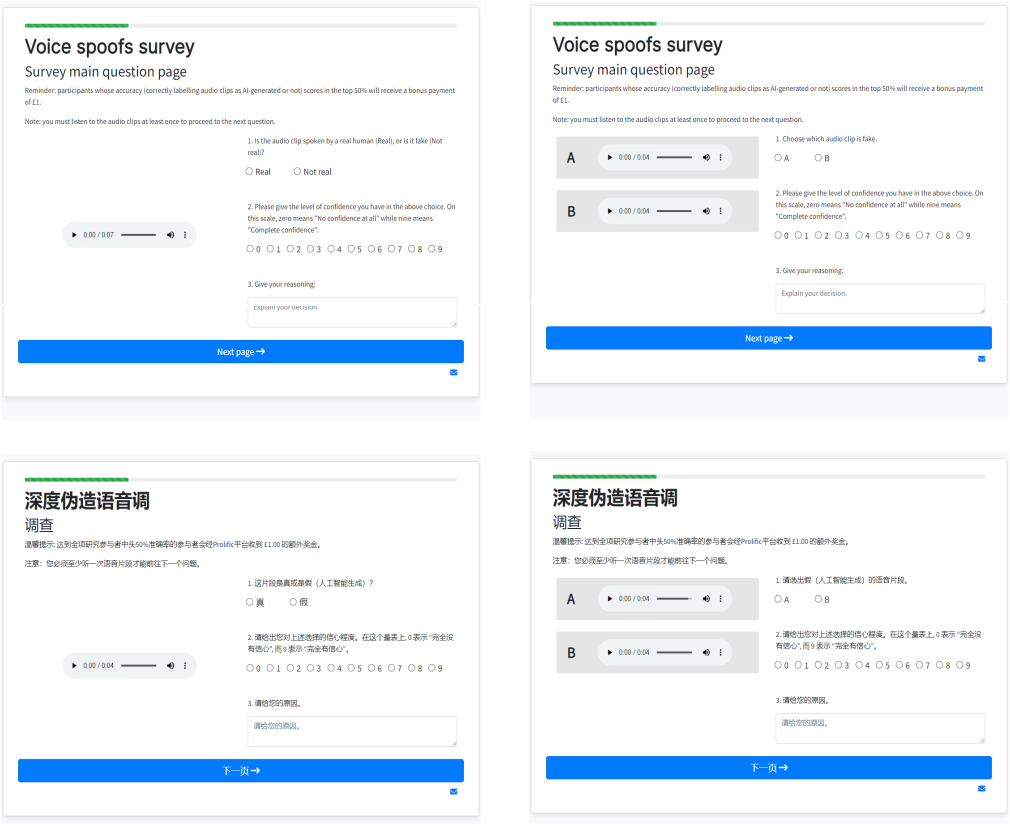

图2. 任务界面的截图。

一元

我们向每位参与者展示了随机选取的20个不同片段，每个片段单独显示在一个页面上。参与者收听的真实语音片段和合成音频片段数量大致相等，但我们没有告知他们这一比例。我们要求参与者判断所听到的片段是真实的还是伪造的。

二元

我们展示了随机选取的 20 对音频片段（标记为“A”和“B”），每对片段都包含同一句口语句子。一段由人类说话者说出，另一段由 VITS 生成。我们随机安排伪造片段和真实片段的顺序，并要求参与者判断哪段是伪造的。我们设置这一情境，是为了了解上下文信息是否有助于检测。

熟悉化处理

除了这两种配置外，我们还将一半参与者随机分配到熟悉化处理组。我们设置这一处理，是为了验证现有文献的结论，并了解人类能否像机器学习模型一样通过训练来检测深度伪造。在开始主要检测任务前，我们向处理组的参与者展示了五段深度伪造话语。我们告知参与者这些示例是合成的，并允许他们多次收听这些片段。这些片段与主要任务中使用的刺激材料不同。

对于对照组的参与者，我们安排了一项填充任务。在这项任务中，我们请参与者列举合成语音的潜在应用，并表达他们对于合成音频会对社会产生正面还是负面影响的看法。

参与者

我们通过Prolific平台招募参与者，并筛选出英语和普通话流利的参与者，因为流利度会影响检测性能 [30]。我们按每小时7.25英镑的标准支付报酬。为了鼓励参与者更认真地作答，我们告知他们，如果检测得分位于前50%，便可获得1英镑奖金。我们总共招募了529名参与者。平均年龄为28.9岁，50.6%的参与者自我认同为男性。表2详细列出了各处理组的人口统计特征。

表2. 各组参与者人数。

组	英语			普通话		
	n	年龄（标准差）	男性（%）	n	年龄（标准差）	男性（%）
单元无熟悉化	76	26.8 (8.1)	55.2	65	31.0 (10.4)	44.6
单元熟悉化	65	26.7 (7.3)	56.9	54	31.4 (8.7)	44.4
二元无熟悉化	60	27.5 (7.2)	53.3	70	31.8 (9.0)	48.8
二元熟悉化	80	27.4 (7.3)	57.5	59	29.1 (8.5)	39.6
总体	281	27.1 (7.5)	55.8	248	30.9 (9.2)	44.5

伦理声明

该研究经伦敦大学学院安全与犯罪科学系伦理委员会审查后获豁免。所有参与者均已获知

研究目的，且年龄均超过18岁。参与前，参与者需勾选一系列复选框，以提供书面知情同意。

与自动深度伪造检测器进行基准比较

为了将人类参与者的表现与自动化方法进行比较，我们训练了两个专门用于检测语音深度伪造的人工神经网络。这两个网络采用了 LFCC-LCNN架构 [41]。LFCC-LCNN 将原始音频波形转换为二维表示。它们通过接触真实样本和深度伪造样本进行学习，并在正确判定样本真伪时获得奖励。ASVspoof 2021挑战赛将 LFCC-LCNN 用作欺骗检测的基线模型 [9]。因此，它们是我们实验中合理的基准。关于表现最佳的语音深度伪造检测架构的更多详情，请参阅总结 ASVspoof 的文章 2021 [9]。

我们为每种语言使用了两个版本：

1. 域内：我们使用 LJSpeech 和 CSMSC 的训练划分中的真实样本训练网络，并将训练划分中的句子输入 VITS，以此生成深度伪造样本。

2. 域外：我们使用 FAD [42], 这一另一个普通话数据集训练普通话网络。我们使用预训练ASVspoof网络 [43] 进行英语评估。

众所周知，人工神经网络在评估与训练期间见过的样本相似的数据时表现良好。然而，遇到不同的示例时，其表现往往会下降 [44], ，即使这些示例使用的是同一种语言。对于人类听众来说，这些差异可能很细微，包括说话人身份或环境的变化。因此，为了更公平地与人类表现进行比较，我们引入了域外版本；尤其是因为本研究的参与者不太可能认出 LJSpeech 和 CSMSC 中的说话人身份。

结果

总体表现

图3总结了人类在所有不同组别中的表现。表3和表4列出了分类选择的细分结果，汇总了英语和普通话的结果。我们使用 SciPy [45] 和 statsmodels [46] Python 软件包完成了分析。更多详情及各刺激材料的结果，请参阅补充信息。

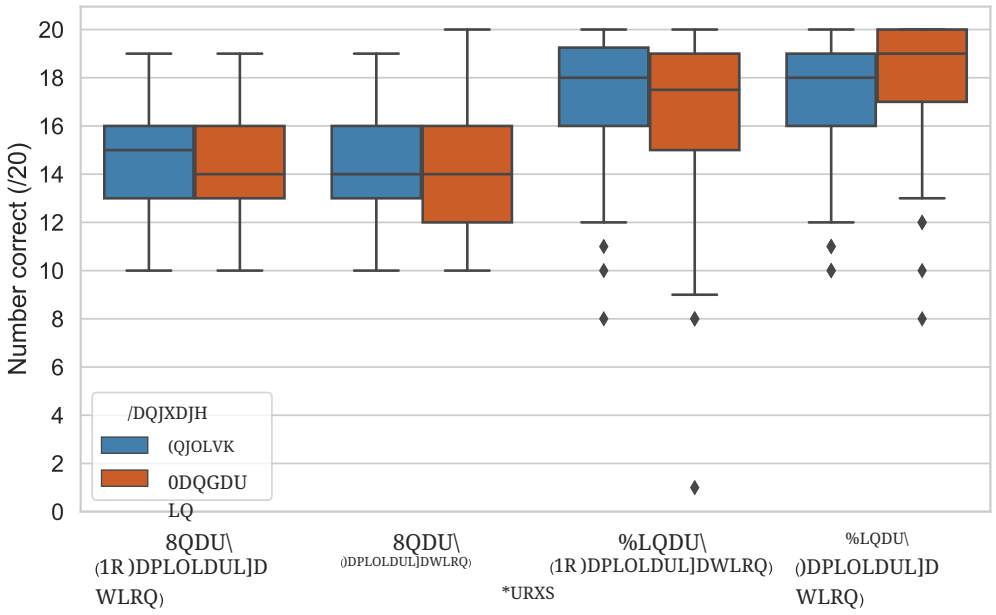

图3. 汇总不同组人类表现的箱线图 .

表3. 单项组回答的混淆矩阵。

		预测类别	
		真实（*2,442*）	伪造（*2,678*）
真实类别	真实（*2,598*）	1,761	837
	伪造（*2,522*）	681	1,841

n = 5,120. 总体准确率 = 70.35%。
正确识别的真实内容 = 67.78%。
正确识别的伪造内容 = 73.0%。

在一元情境中，参与者有 70.35% 的时间做出了正确分类。他们更善于识别深度伪造内容（准确率为 73%）。相比之下，

表 4. 二元组响应的混淆矩阵。

		预测类别	
		Real	*Fake*
真实类别	*Real*	-	-
	伪造（*5,380*）	775	4,605

在此情境中，真实真实类别标签未定义，因为参与者被要求每次都选择伪造片段。总体准确率相当于正确识别伪造内容的比例，为 $= 85.59\%$。

参与者有 67.78% 的时间正确识别出真实语音示例。我们推测，真实样本被误分类的比例较高，部分原因是参与者的怀疑态度有所增强，因为任务说明让他们知道存在深度伪造内容。这一行为与 Bray 等人 [22] 的观测结果一致。

在二元场景下，表现有所提升。参与者在 85.59% 的试验中正确识别出深度伪造音频。然而，二元设置并不符合现实情况。即使已知说话人身份，也不太可能获得包含与待评估测试片段相同语音内容的参考语句。

衡量干预措施的效果

我们采用与 Groh 等人类似的方法 [24]，以分离各项干预措施对表现的影响。我们根据参与者提供的置信度评分对其决策进行加权，将正确／错误结果转换为连续值。

参与者完成的十分制置信度量表用作映射函数。最低分 0 表示参与者是在猜测，因此其对作出正确决策的信心程度相当于 50%。相反，最高分 9 则相当于 100% 的确信程度。

最终转换后的分数取决于参与者是否做出了正确分类。例如，如果参与者将自己的信心程度评为 7，这对应着 88% 的置信度。如果他们做出了正确决策，调整后的分数就是 0.88。反之，如果他们做出了错误决策，我们就用 1 减去该值，得到 0.12 的调整后分数。

修订后的分数也能让我们与自动深度伪造检测器进行更公平的比较；这些检测器在评估示例时会输出介于 0 和 1 之间的分数。在本文余下部分，我们将修订后的分数称为准确率分数。我们还会将分数重新缩放为百分比。

对结果进行转换后，我们使用线性回归分析了不同干预措施对参与者在每个音频片段上的准确率分数的影响。除语言、熟悉化和二元干预外，我们还分析了片段时长的影响。表5概述了总体、一元和二元层面的结果。

表5：干预措施对置信度加权准确率的线性回归结果。

自变量	因变量：置信度加权准确率		
	全部（标准差）	一元（标准差）	二元（标准差）
常数	43.742***（1.897）	46.394***（2.791）	71.217***（2.530）
普通话[a]	1.790（1.404）	1.477 (1.882)	2.152 (2.118)
熟悉化	3.840*** (1.191)	3.758** (1.571)	3.854** (1.802)
片段长度	0.797*** (0.209)	0.375 (0.358)	1.168*** (0.230)
二元干预	29.830*** (1.186)	-	-
观测值	10,500	5,120	5,380
R^2	0.165	0.003	0.010
调整后的 R^2	0.165	0.002	0.009
F 统计量	171.102***	2.455*	11.794***

[a]指示任务中使用了哪种语言的虚拟变量。 $1 =$ 普通话， $0 =$英语。$^{*}p < 0.1$。$^{**}p < 0.05$。$^{***}p < 0.01$。

参考音频有助于检测深度伪造

线性回归结果表明，二元场景带来的提升在统计学上显著（$p < 0.001$）。因此，结果显示，通过参考音频提供上下文信息，有助于揭示合成语音的特征。

训练人类检测深度伪造只能略微提升表现

熟悉化处理平均可将检测准确率提高 3.84%（$p = 0.001$）。一元回归和二元回归结果也显示了这一效果，准确率分别提高 3.76%（$p = 0.017$）和 3.85%（$p = 0.032$）。然而，在其他条件不变的情况下，即使纳入熟悉化处理，对于平均片段长度（5.76 秒）的片段，一元情境下的准确率也仅略高于随机水平（52.31%）。

检测普通话和英语中的深度伪造同样困难

图3显示，在不同处理组中，英语和普通话的表现相近。表5也支持这一观察：普通话使用者的表现仅比英语使用者高出 1.79%，且这一差异在统计上并不显著（$p = 0.202$）。

较短的语音深度伪造并不更容易识别

由于我们的刺激材料时长介于 2 到 11 秒之间，我们在回归分析中纳入了片段长度，以验证较短的片段是否更容易辨别。结果表明，片段长度对准确率的影响微乎其微，每增加一秒，表现仅提高 0.80%。我们的散点图（图4）支持这一结论，并显示这两个变量之间没有关联。这些发现与 Watson 等人 [28],的观点相矛盾；他们认为较短的深度伪造更容易识别。



图4。散点图显示片段长度与置信度加权准确率之间的关系。

分析表现与时间的关系

除了分析处理效应外，我们还考察了在任务上花费更多时间这一假设是否能提升表现。

更频繁地听这些片段并不能帮助检测

我们记录了参与者点击每个音频片段的次数，并将其与准确率进行比较。如图5所示，这两个变量之间没有关联（$\rho = -0.05, p < 0.001$）。

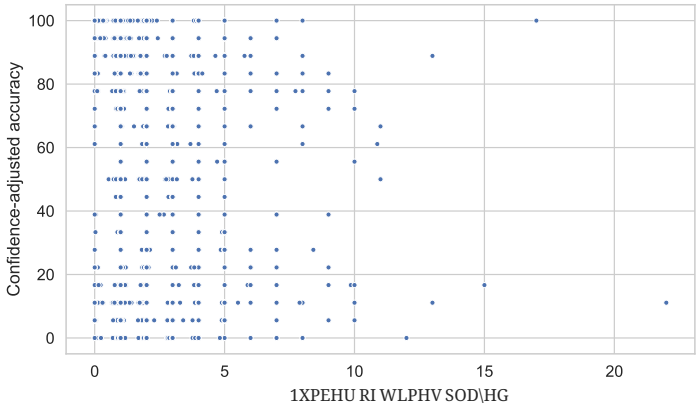

图5。散点图显示播放次数与置信度加权准确率之间的关系。

在任务上花费更多时间也不会影响表现

与上述分析类似，我们将完成整个任务所花费的时间与正确识别的片段总数进行了比较。图6显示这两个变量之间没有关联（$\rho = 0.10$, $p = 0.018$），这表明花更多时间完成任务并不能提升表现。

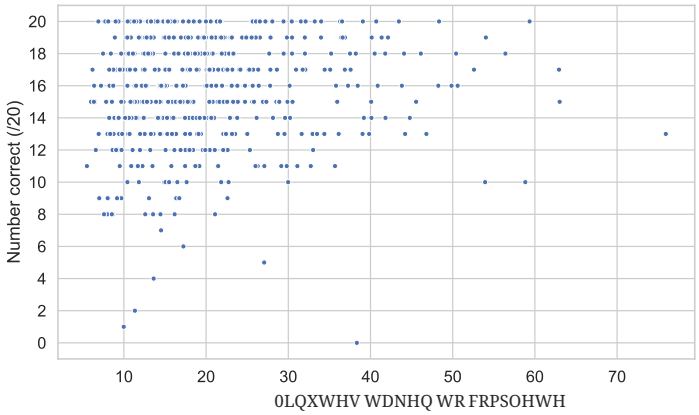

图6. 散点图显示完成任务所用时间（分钟）与正确性得分之间的关系。

没有明确反馈，参与者在整个任务过程中不会有所进步

为了了解参与者是否会随着看到更多示例、在任务中推进得更深入而有所进步，我们计算了每个问题编号对应的正确回答数量。如果确实如此，那么第 20 题的正确回答数应多于第 1 题。图 7 展示了由此得出的直方图。直方图显示，各题的表现相对稳定。这一观察表明，除非获得明确反馈，否则参与者在整个任务过程中并不会进步；Groh 等人 [24] 和 M¨uller 等人 [30]的研究也考察了这一点。我们通过针对均匀分布进行单向卡方假设检验，对这一结果进行了定量验证，结果未达到统计显著性（$\chi^2 = 6.19, \quad p = 0.997$）。

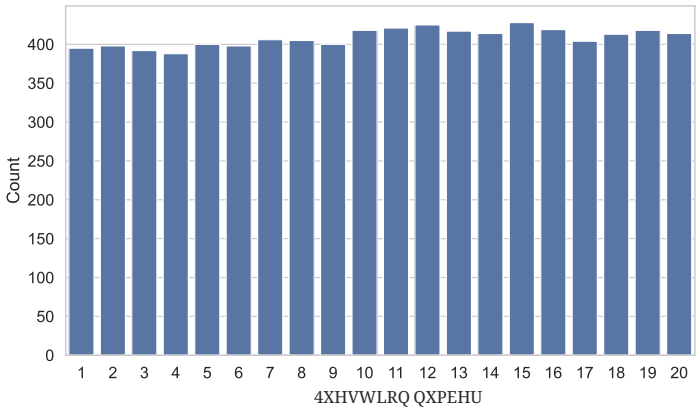

图 7. 各问题编号对应的正确回答直方图。

比较人类表现与自动检测器

下一节将比较人类表现与自动深度伪造检测器。为便于比较，我们采用机器学习文献中常见的性能指标。

- **受试者工作特征（ROC）**：这些图表示判别能力。它们比较不同阈值下的真阳性率与假阳性率。

- **受试者工作特征曲线下面积（AUROC）**：该分数将 ROC 汇总为一个数值。AUROC 为 50% 表示所有预测都是猜测，而 AUROC 为 100% 则表示在所有试验中都能完美区分真实样本与深度伪造。

- **等错误率（EER）**：这描述了 ROC 曲线上真阳性率与假阳性率相等的点。

图 8 展示了 AUROC 和等错误率（EER）得分。由于人类与自动检测器的推理设置更具可比性，我们在此分析中仅纳入一元情境。两者每次都评估一个片段。由于观察到英语和普通话的结果相似，我们将两者的结果汇总。

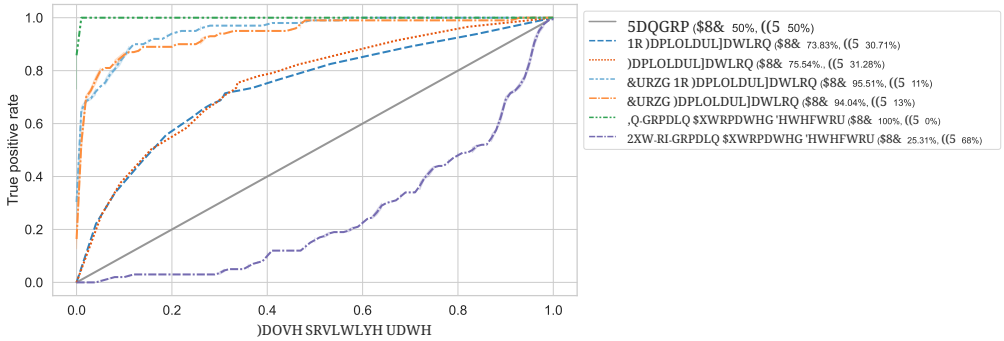

图 8. 一元情境下的受试者工作特征曲线。

与自动检测器相比，人类表现受未知条件的影响较小

无熟悉化曲线（AUROC ＝ 73.83%）和熟悉化曲线（AUROC ＝75.54%）证实，人类的表现优于随机猜测。这些曲线也支持线性回归的结果。向参与者展示深度伪造示例对表现的影响微乎其微。然而，表现相当不稳定：平均而言，人类有四分之一的时间会错误地对片段进行分类。

人类的表现不如域内自动检测器，这些检测器具有完美的区分能力（两种语言的 AUROC 均为 ＝ 100%）。然而，域外检测器经常将真实样本错误地分类为深度伪造（AUROC ＝ 25.31%）。根据这一表现，人类对说话人身份等未知因素更具鲁棒性。

群体语音深度伪造检测的表现可与顶尖自动检测器相媲美

根据 Groh 等人 [24], ，我们对参与者针对每个片段的准确率分数取平均值，以计算众包响应。与视频刺激材料 [24], 的结果相似，群体表现与领域内检测器不相上下。然而，对参与者进行熟悉化训练带来的益处在对响应取平均值后便消失了。群体无熟悉化和群体熟悉化的 AUROC 值相近，分别为 95.51% 和 94.04%。

自由文本分析

为了了解参与者如何判断音频片段的真实性，我们分析了他们的自由文本回答。我们按语言、片段真实性以及

参与者是否做出了正确选择对回答进行了分组。随后，我们使用词频-逆文档频率权重创建了词云。词频-逆文档频率用于衡量一个词在某份文档中的重要性，并与整个文档集进行比较，以考虑频繁出现的词语 [47]。图 9 和图 10 展示了英语和普通话词云。

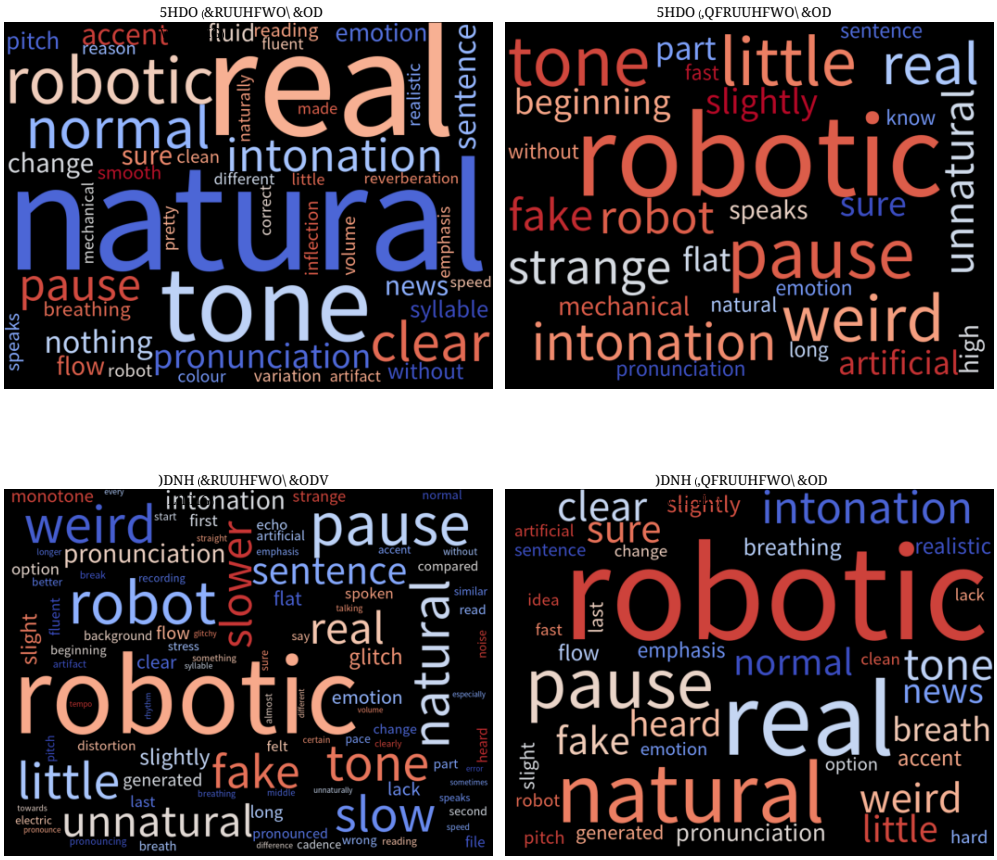

图 9. 英语片段理由词云。

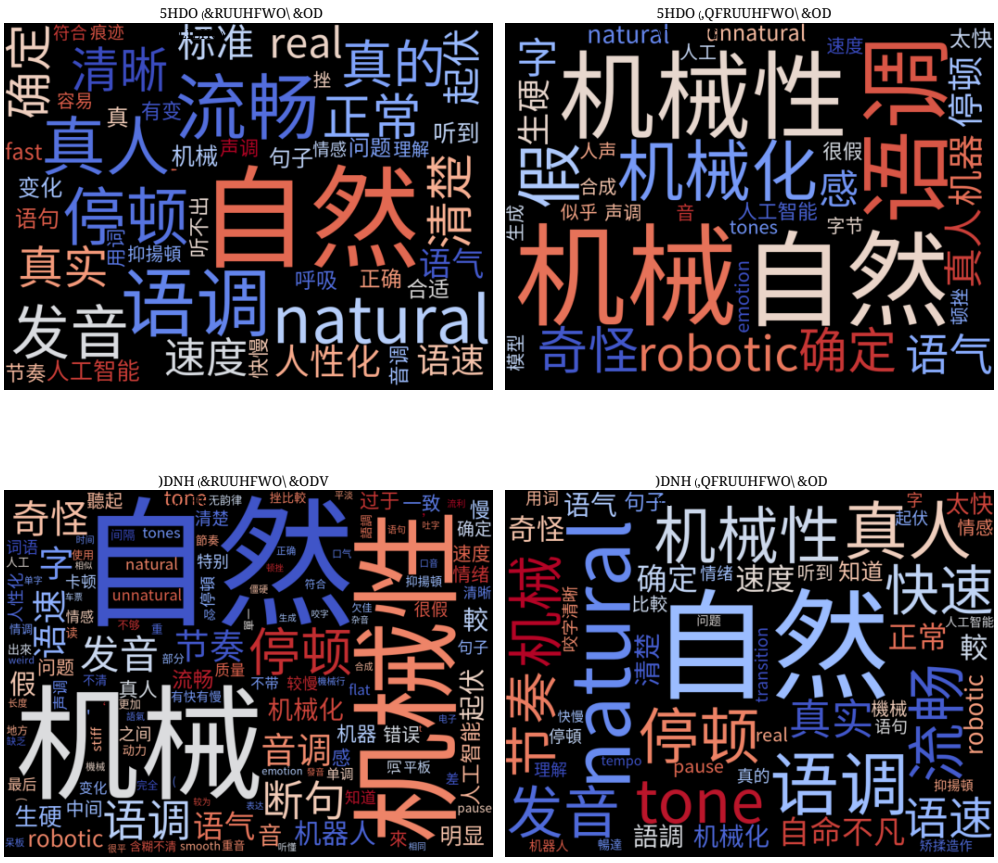

图10. 普通话片段的理由词云。请注意，普通话任务的参与者提供的理由使用了普通话和英语两种语言。

无论是否做出了正确决策，参与者提及的特征都相同。例如，在图 9 中，正确将真实语音分类为真实语音的参与者（图 9 左上方）提到了停顿、语气和语调。然而，错误地将真实语音归类为伪造语音的参与者（图 9 右上方）也提到了这些相同的特征。我们根据片段的实际标签以及参与者是否作出了正确判断，对回答进行了比较。我们没有发现这些组别之间存在显著差异。因此，整合这些人类特征的自动检测器只能带来有限的改进。我们在英语和普通话中都观察到了这一现象。参与者往往依靠直觉进行分类，提及自然度（自然）和机械感（机械）等声音特征。除了直觉之外，英语和普通话使用者也经常提到停顿（停顿）、语调（语调）、发音（发音）和速度（速度）。

关于语言之间的差异，英语使用者更多地提到了呼吸。相比之下，普通话使用者提到了说话者的节奏（节奏）、词间停顿（断句）和流利度（流 畅）。这一结果可能是由这两种语言在时序特性上的差异造成的。英语是重音计时语言，而普通话是音节计时语言 [48]。

局限性

尽管我们的实验设置能够与自动检测器进行比较，但不一定能反映听众可能遇到语音深度伪造的更真实场景。

首先，我们在实验中呈现的深度伪造比例并不能反映现实环境中的实际比例。在这项任务中，参与者遇到深度伪造和真实样本的可能性相同。然而，人工智能生成内容（包括出于恶意目的使用深度伪造）目前仍然罕见。此外，由于我们向参与者说明了任务的性质，他们可能会对深度伪造的出现更加留意。

此外，我们尽量减少了刺激材料中的上下文信息。例如，我们没有考察听众的语境知识（例如对说话人身份、情绪状态、对话参与方数量或政治隶属关系的了解）可能影响其决策的情形。这些方面可能与语音深度伪造可能出现的典型应用场景相关，例如虚假新闻传播 [5]。未来的研究可以探讨这些特征如何影响检测。

此外，我们请两种语言的参与者聆听声称来自同一位女性说话人的话语。鉴于说话人的年龄和性别会影响言语感知， [49, 50], 未来的研究可以探讨说话人身份的变化如何影响深度伪造检测性能。

为了生成我们的深度伪造刺激材料，我们采用了一种较旧的方法，它未必能代表最先进水平的语音合成算法。尽管我们的结果表明，人类能够在多大程度上检测出利用有限计算资源生成的语音深度伪造，但这些结果未必能准确反映当前最先进条件下的表现。

讨论

人类能够检测出语音深度伪造，但无法始终如一地做到这一点。无论语言如何，人们往往依靠自然度来识别深度伪造。随着语音合成算法不断改进、生成的语音越来越自然，人类将更难识别语音深度伪造。

尽管英语和普通话说话人在用于检测深度伪造的特征方面存在一些差异，但这两个组也有许多相似之处。因此，无论涉及何种语言，语音深度伪造造成的威胁程度都相同。

随着合成数据的计算门槛降低，对手将更容易生成更多深度伪造内容。更多出现在现实环境中的深度伪造内容会产生连锁影响。对手将有更多机会扩大行动规模，尤其是利用冒充和鱼叉式网络钓鱼等方式散布虚假信息 [6]。

归根结底，深度伪造内容的制作与检测之间是一场军备竞赛 [51]。我们该如何避免落入深度伪造的骗局？我们的二元场景表明，如果我们知道说话人身份，将音频与参考音频进行比较会有所帮助。然而，我们并不总是掌握这些信息。

正如我们的熟悉化结果所示，通过向人们展示深度伪造音频示例来提高意识，效果有限。花更多时间评估这些片段似乎也没有帮助。

总而言之，试图提升人类的检测能力并不现实。我们的研究表明，即使在受控环境中，任务相对容易（参与者知道存在语音深度伪造，且这些深度伪造并非由最先进水平的语音合成器生成），深度伪造检测的表现仍不理想。我们的结果表明，需要借助自动检测器来弥补人类听众的弱点。自动检测器在域内数据上的表现说明，它们能够识别出人类无法察觉的细微差别。不过，我们也发现，这些检测器很脆弱，一旦测试音频的环境条件发生变化，便无法正常工作。鉴于人类能力的局限，以及可用于部署检测器的计算资源日益充足，研究应着重改进这些检测器。在此期间，众包是合理的缓解措施。我们证实，群体表现可与表现最佳的自动检测器相媲美，而且不那么脆弱。将事实核查

工具扩展至纳入音频评估，是防范深度伪造威胁的一种方式。

补充信息

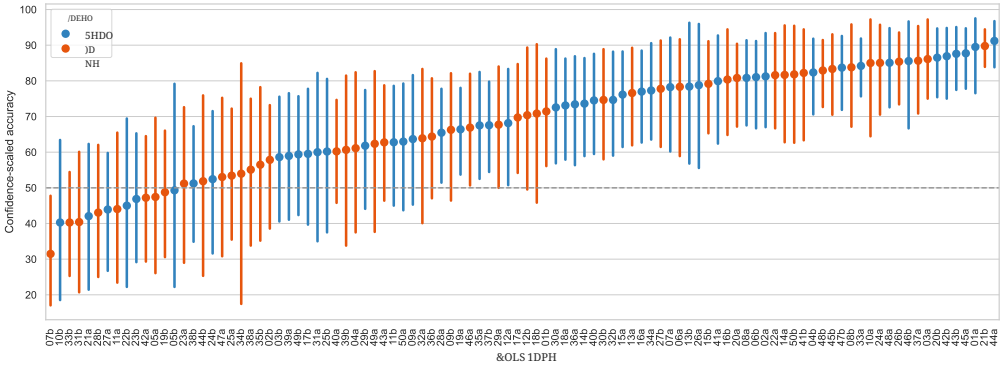

图 S1. 每个片段的置信度调整后的准确率得分（英语，一元，无熟悉化）。

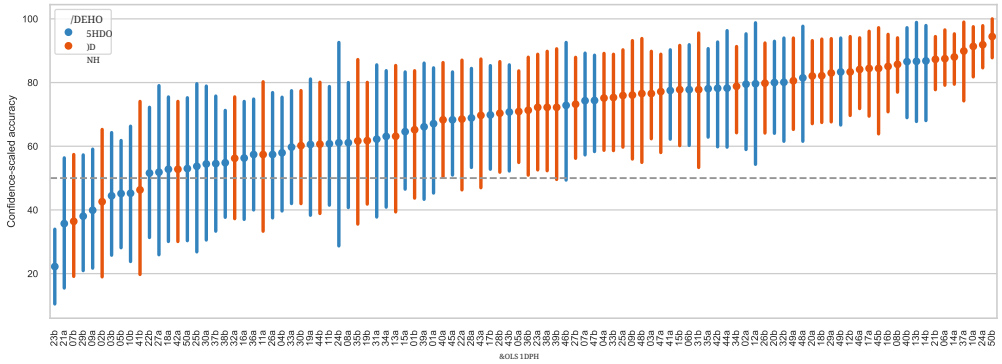

图 S2. 每个片段的置信度调整后的准确率得分（英语，一元，熟悉化）

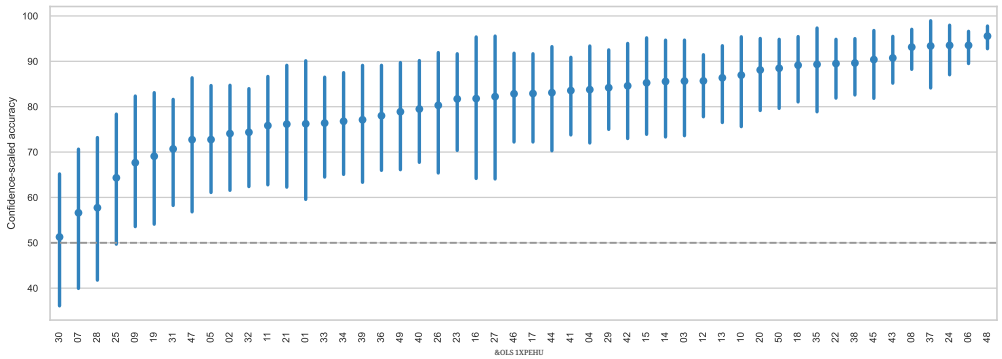

图 S3. 每个片段的置信度调整后的准确率得分（英语、二元、无熟悉化）。

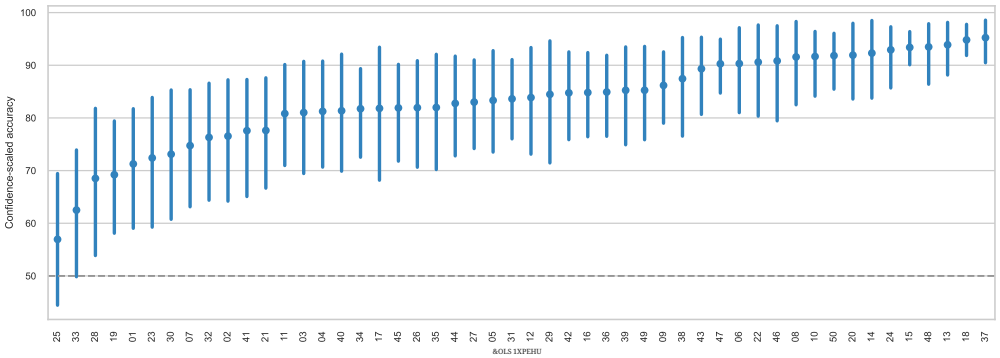

图 S4. 每个片段的置信度调整后的准确率得分（英语，二元，熟悉化）

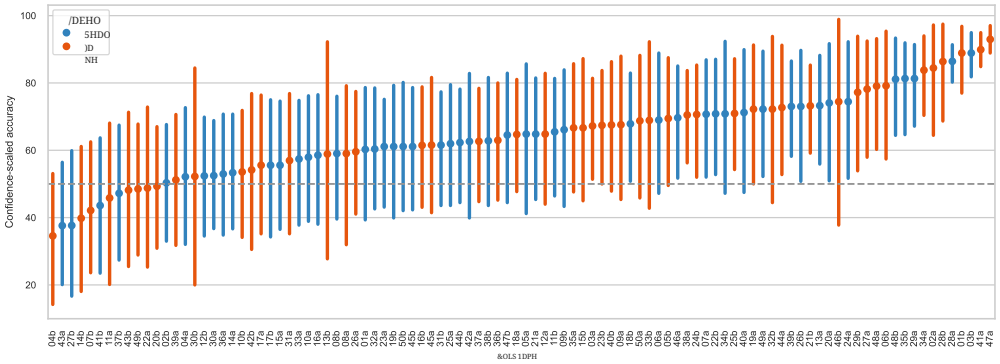

图 S5. 每个片段的置信度调整后的准确率得分（普通话，一元，无熟悉化）。

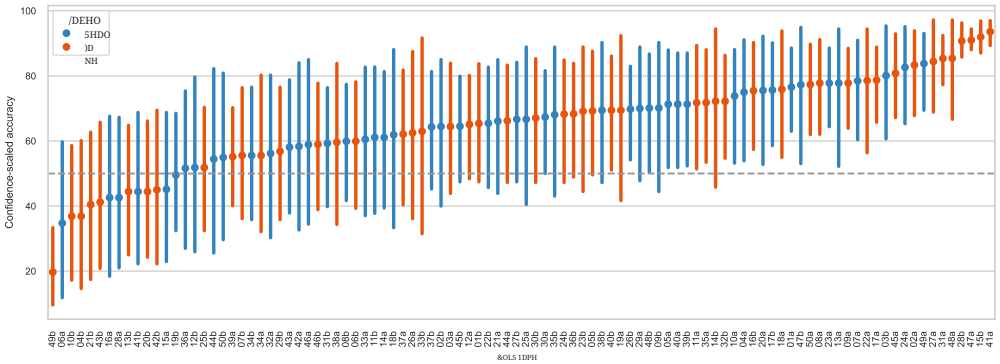

图 S6. 每个片段的置信度调整后的准确率得分（普通话，一元，熟悉化）。

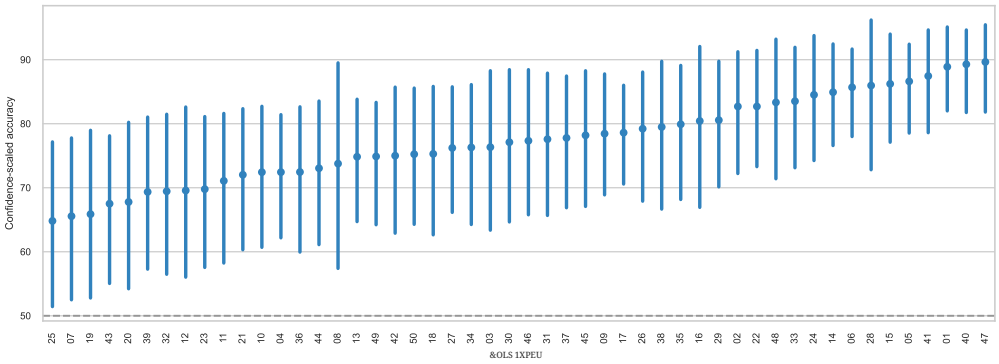

图 S7. 普通话二元无熟悉化条件下各片段的置信度调整后的准确率得分。

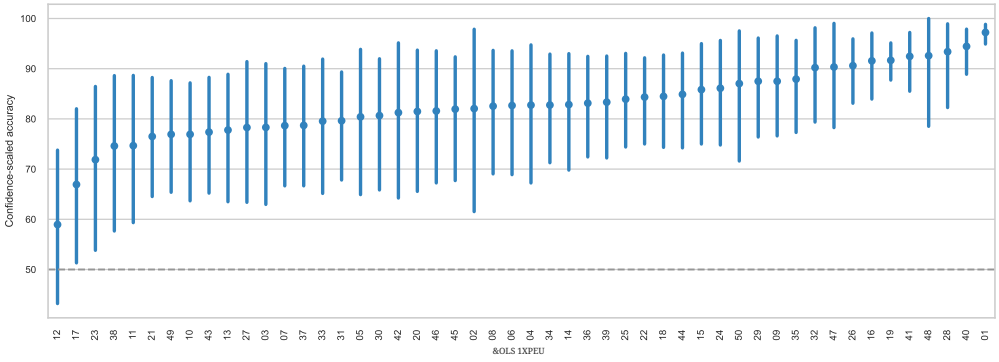

图 S8. 普通话二元熟悉化条件下各片段的置信度调整后的准确率得分。

致谢

Kimberly T. Mai 感谢 Kelvin Ma 协助翻译调查材料 .

参考文献

1. Brewster T. 诈骗分子克隆公司董事的声音，实施了价值 3500 万美元的银行劫案，警方发现；2022年。来源：https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=7dfbccf67559。2. Stupp C. 诈骗分子利用人工智能模仿首席执行官的声音，犯下罕见的网络犯罪；2019年。来源：https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-115

67157

3. Choi S、Han S、Kim D、Ha S。Attentron：利用基于注意力的可变长度嵌入的少样本文本转语音。载于：Meng H、Xu B、Zheng TF 编。国际语音通信协会第21届年会 Interspeech 2020，线上活动，中国上海，2020年10月25日至29日。ISCA；2020年。第 2007–2011 页。来源：https://doi.org/10.21437/Interspeech.2020-2096。

4. Alspach K. 你的老板听起来有点古怪吗？这可能是音频深度伪造；2022年。来源：https: //www.protocol.com/enterprise/deepfake-voice-cyberattack-ai-audio. 5. Caldwell M, Andrews J, Tanay T, Griffin L. 人工智能赋能的未来犯罪。《犯罪科学》。2020年；9(1):1–13. 6. Mirsky Y, Demontis A, Kotak J, Shankar R, Gelei D, Yang L, et al. 攻击性人工智能对组织构成的威胁。《计算机与安全》。2022年；第103006页。7. Wu Z, Yamagishi J, Kinnunen T, Hanil̨ci C, Sahidullah M, Sizov A, et al. ASVspoof：自动说话人验证欺骗与对策挑战赛。《IEEE 信号处理专题精选期刊》。2017年；11(4):588–604. doi:10.1109/JSTSP.2017.2671435.

8. Nautsch A, Wang X, Evans NWD, Kinnunen TH, Vestman V, Todisco M, et al. ASVspoof 2019：用于检测合成、转换和重放语音的欺骗对策。《IEEE 生物医学与行为身份科学汇刊》。2021年；3(2):252–265. doi:10.1109/TBIOM.2021.3059479.

9. Yamagishi J, Wang X, Todisco M, Sahidullah M, Patino J, Nautsch A 等。ASVspoof 2021：加快伪造语音和深度伪造语音检测的进展。arXiv 预印本 arXiv:210900537。2021年。

10. Gamage D, Chen J, Ghasiya P, Sasahara K。《深度伪造与社会：未来将走向何方？》载于：《虚假媒体生成与检测前沿》。Springer；2022年。第3–43页。

11. Goodfellow I, Bengio Y, Courville A。《深度学习》。MIT Press；2016年。

12. Zhang Y, Tiˇno P, Leonardis A, Tang K。神经网络可解释性综述。《IEEE 新兴计算智能主题汇刊》。2021年。

13. van Huijstee M, van Boheemen P, Das D, Nierling L, Jahnel J, Karaboga M 等。《应对欧洲政策中的深度伪造》。欧洲议会；2021年。可访问：https://books.google.co.uk/books?id=O8SrzgEAC。

14. Schick N.《深度伪造与信息末日：你亟须了解的内容》。Hachette；2020年。获取自：https://books.google.co.uk/books?id=K4HTDwAAQBAJ。

15. Goodfellow I、Pouget-Abadie J、Mirza M、Xu B、Warde-Farley D、Ozair S 等。生成对抗网络。载于：Ghahramani Z、Welling M、Cortes C、Lawrence N、Weinberger KQ 编。《神经信息处理系统进展》第27卷。Curran Associates, Inc.；2014年。获取自：https://proceedings.neurips.cc/paper_files/paper/2014/file/ 5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf。

16. Sohl-Dickstein J、Weiss E、Maheswaranathan N、Ganguli S. 使用非平衡热力学进行深度无监督学习。载于：Bach F、Blei D，编者。《第32届国际机器学习会议论文集》第37卷

《机器学习研究论文集》。法国里尔：PMLR；2015年。第2256–2265页。网址：https://proceedings.mlr.press/v37/sohl-dickstein15.html。

17. Damer N, Saladi´e AM, Braun A, Kuijper A. MorGAN：生成对抗网络创建的人脸变形攻击的识别脆弱性与攻击可检测性。载于：2018年 IEEE 第九届生物特征理论、应用与系统国际会议（BTAS）；2018年。第 1–10 页。

18. Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar SK. 人脸替换：自动替换照片中的人脸。ACM图形学汇刊。2008年；27(3)：1–8。doi:10.1145/1360612.1360638。

19. Zhang T. 深度伪造生成与检测：综述。多媒体工具与应用。2022年；81(5)：6259–6276。doi:10.1007/s11042-021-11733-y。

20. Barari S, Lucas C, Munger K. 政治深度伪造与其他虚假媒体及（有时）真实媒体同样可信；2021年。可访问：osf.io/cdfh3。

21. Appel M, Prietzel F. 政治深度伪造检测. 计算机媒介传播杂志. 2022年; 27(4). doi:10.1093/jcmc/zmac008.

22. Bray SD, Johnson SD, Kleinberg B. 测试人类检测人脸深度伪造图像的能力. arXiv预印本 arXiv:221205056. 2022年;.

23. Nightingale SJ, Farid H. 人工智能合成人脸与真实人脸难以区分，且更值得信任. 美国国家科学院院刊. 2022年;119(8):e2120481119. doi:10.1073/pnas.2120481119.

24. Groh M, Epstein Z, Firestone C, Picard R. 人类群体、机器和机器辅助群体进行深度伪造检测. 美国国家科学院院刊. 2022年;119(1):e2110013119. doi:10.1073/pnas.2110013119.

25. K¨obis NC、Dole˘zalov´a B、Soraperra I。双重受骗：人们无法识别深度伪造，却认为自己能识别。iScience。2021年；24(11)：103364。doi：https://doi.org/10.1016/j.isci.2021.103364。

26. Tahir R、Batool B、Jamshed H、Jameel M、Anwar M、Ahmed F 等。眼见为实：探索深度伪造视频中的感知差异。载于：2021年计算机系统中的人为因素CHI会议论文集；2021年。第1–16页。

27. Geirhos R, Zimmermann RS, Bilodeau BL, Brendel W, Kim B. 别相信你的眼睛：特征可视化的（不）可靠性。CoRR. 2023年；abs/2306.04719. doi:10.48550/arXiv.2306.04719。

28. Watson G, Khanjani Z, Janeja VP. 大学生群体中的音频深度伪造感知。arXiv 预印本 arXiv:211203351。2021年；。

29. Wang X, Yamagishi J, Todisco M, Delgado H, Nautsch A, Evans N, 等。ASVspoof 2019：合成、转换和重放语音的大型公共数据库。《计算机语音与语言》。2020年；64:101114。

30. M¨uller NM, Pizzi K, Williams J. 人类对音频深度伪造的感知。载于：第一届音频多媒体深度伪造检测国际研讨会论文集；2022年。第 85–91 页。

31. Karras T、Laine S、Aittala M、Hellsten J、Lehtinen J、Aila T。分析并提升 StyleGAN 的图像质量。载于：IEEE/CVF计算机视觉与模式识别会议论文集；2019。第 8110–8119 页。

32. Dolhansky B、Bitton J、Pflaum B、Lu J、Howes R、Wang M 等。深度伪造检测挑战数据集。arXiv 预印本 arXiv:200607397。2019年。

33. Li Y、Yang X、Sun P、Qi H、Lyu S。Celeb-DF：用于深度伪造取证的大规模挑战性数据集。载于：IEEE/CVF计算机视觉与模式识别会议论文集；2019年。第 3207–3216 页。

34. Rossler A、Cozzolino D、Verdoliva L、Riess C、Thies J、Nießner M。Faceforensics++：学习检测篡改的人脸图像。载于：IEEE/CVF国际计算机视觉会议论文集；2019年。第 1–11 页。

35. Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Um´e C 等。DeepFaceLab：集成式、灵活且可扩展的换脸框架。arXiv 预印本 arXiv:200505535。2020年;。

36. Kumar K, Kumar R, De Boissiere T, Gestin L, Teoh WZ, Sotelo J 等。Melgan：用于条件波形合成的生成对抗网络。神经信息处理系统进展。2019年;32。

37. Ito K, Johnson L. LJ语音数据集；2017年。https://keithito.com/LJ-Speech-Dataset/。

38. Databaker. 中文标准普通话语料库；2019年。https://www.data-baker.com/open_source.html。

39. Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y 等。ESPnet：端到端语音处理工具包。载于：Interspeech 会议论文集；2018年。第2207–2211页。获取地址：http://dx.doi.org/10.21437/Interspeech.2018-1456。

40. Kim J, Kong J, Son J. 用于端到端文本转语音的对抗学习条件变分自编码器。载于：国际机器学习会议。PMLR；2021年。第5530–5540页。

41. Wang X, Yamagishi J. 面向合成语音检测的近期神经网络欺骗对策比较研究。载于：Hermansky H, Cernock´y H, Burget L, Lamel L, Scharenborg O, Motl´ıcek P 编。Interspeech 2021，国际语音通信协会第22届年会，捷克布尔诺，2021年8月30日至9月3日。ISCA；2021年。第4259–4263页。获取地址：https://doi.org/10.21437/Interspeech.2021-702。

42. Ma H、Yi J、Wang C、Yan X、Tao J、Wang T 等。FAD：中文伪造音频检测数据集。arXiv 预印本 arXiv:220712308。2022年。

43. Delgado H、Evans N、Kinnunen T、Lee KA、Liu X、Nautsch A 等。ASVspoof 2021：自动说话人验证欺骗与对策挑战赛评估计划。arXiv 预印本 arXiv:210900535。2021年。

44. Shen Z, Liu J, He Y, Zhang X, Xu R, Yu H 等. 面向分布外泛化：一篇综述。arXiv 预印本 arXiv:210813624. 2021年;.

45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D 等. SciPy 1.0：Python 科学计算的基础算法。《自然·方法》。2020年;17:261–272. doi:10.1038/s41592-019-0686-2.

46. Seabold S, Perktold J. statsmodels：使用 Python 进行计量经济学与统计建模。载于：第九届 Python 科学会议；2010。

47. Leskovec J, Rajaraman A, Ullman JD.《海量数据挖掘》。第2版。美国：剑桥大学出版社；2014。

48. Mok PPK, Dellwo V. 使用声学节律指标比较母语与非母语语音节律：粤语、北京普通话和英语。载于：Speech Prosody 2008 会议论文集；2008。第423–426页。

49. Hummert ML, Shaner JL, Garstka TA, Henry C. 与老年人沟通：年龄刻板印象、情境和沟通者年龄的影响。《人类传播研究》。1998；25(1)：124–151。

50. Strand EA. 揭示性别刻板印象在言语感知中的作用。《语言与社会心理学杂志》。1999；18(1)：86–100。

51. Chesney R, Citron D. 深度伪造与新型虚假信息战争：后真相地缘政治时代即将来临。《外交事务》。2019年；98：147。