深度学习重塑助听器

助听器佩戴者终于能在嘈杂的房间中辨别声音了。

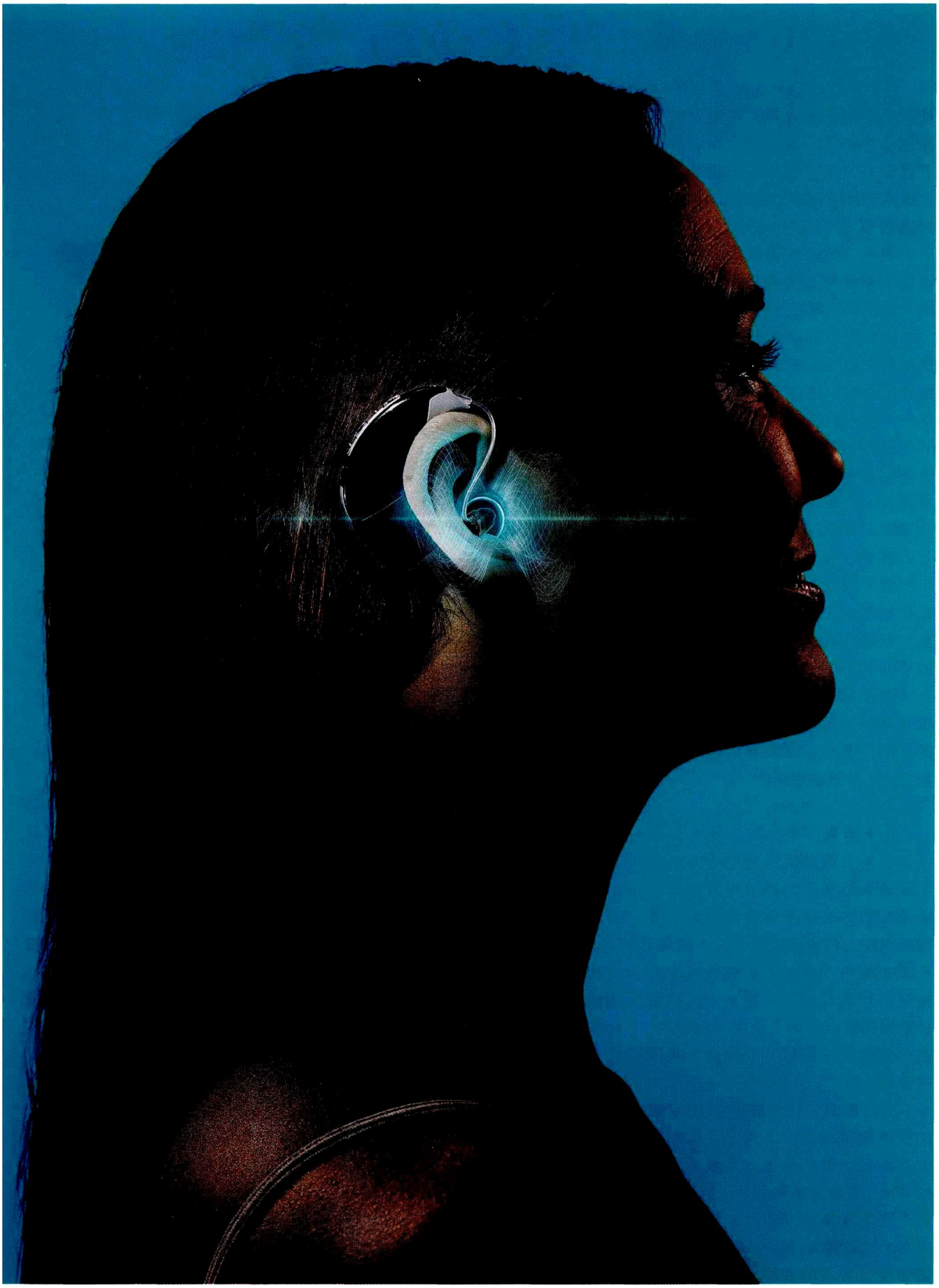
我离家读大学时，我母亲的听力开始下降。每当我回家与她分享我在大学学到的知识时，她都要靠过来才听得清。情况很快变得更糟，如果几个人同时说话，她就无法听清了。现在，即便她佩戴了助听器，也很难区分每个人的声音。我们共进晚餐时，仍然要轮流说话她才能听清楚。

我母亲的听力难题是助听器制造商面临的一个经典问题。人类的听觉系统能够本能、自然地在喧扰的房间中区分不同的声音，但是如何制造出具有这种能力的助听器，已成为数十年来困扰信号处理专家、人工智能专家和听力学家的难题。英国认知科学家科林·奇瑞（Colin Cherry）在1953年首次将这种情况称为“鸡尾酒会问题”。

60多年后，在所有需要助听器的人之中，真正使用助听器的人不到25%。潜在用户感觉最失望的是，助听器无法分辨不同声音。举例来说，当说话声和一辆汽车经过的声音同时出现时，助听器只是将它们的音量调大，生成乱七八糟的喧闹声。

现在是我们解决这个问题的时候了。为了给助听器佩戴者提供更好的体验，我近期在俄亥俄州立大学哥伦布分校的实验室里做了一项实验，将基于深度神经网络的机器学习技术用于分离声音。我们对各个版本的数字过滤器进行测试，这些数字过滤器不仅可以放大声音，还可以从背景噪声中分离出语音并自动调整每个声音的音量。

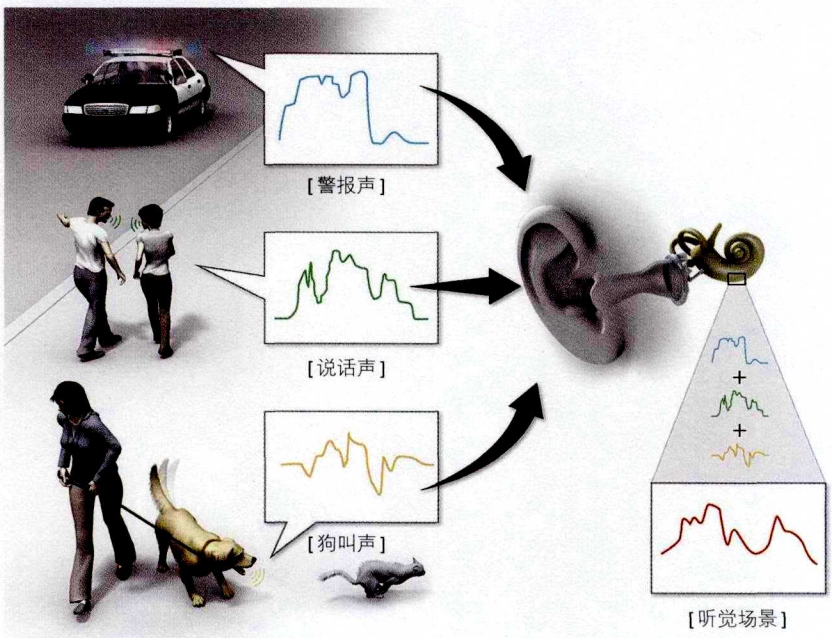
作者：汪德亮



我们相信，这种方法最终能够使听力受损者的理解力恢复到甚至超越正常水平。事实上，我们早期的模型之一就已提高了一些受试者在噪声干扰下理解口语词汇的能力，提高程度从 10% 到 90% 不等。由于听者无须听清短语中的每一个单词即可掌握短语的意思，所以这里的改进通常是指对一句话的理解。

没有更好的助听器，全世界人民的听力水平会变得越来越糟。据世界卫生组织估计，全球 15% 的成年人——大约 7.66 亿人——都存在听力受损的现象。随着人口增长、老龄化比例不断扩大，这一数字还在增加。此外，高级助听器的潜在市场不只局限于听力受损人群。开发人员可以利用该技术完善智能手机的语音识别功能；雇主可以用它来帮助嘈杂工厂车间内的工人；军队可以借此帮助士兵们在混乱喧嚣的战火中听见彼此的声音。

要满足以上新应用的需求，意味着要找到解决“鸡尾酒会问题”的出路。最终，深度神经网络为我们指明了前进的道路。



喧闹的世界：感谢人耳奇怪的形状，使其可以同时捕获不同的声音流。声音流是指来自同一个声源（例如一只狗）的所有声波。这些声音流汇合在一起，就形成了听觉场景（狗叫声 + 警报声 + 说话声）。

几十年来，电气和计算机工程师屡次尝试通过信号处理的方法来实现语音分离，但都以失败告终。最流行的方法是使用语音活动检测器来识别人们说话时话语间的间隙。在该方法中，系统将那些在间隙内捕获的声音认定为“噪声”。然后计算程序会从原始记录中去除噪声，在理想情况下，剩下的就是无噪声的语音。

不幸的是，这种被称为谱减法的技术屡遭诟病，因为它不是删除太多语音，就是保留太多噪声。它频繁地出现令人不快的处理结果（被称为音乐噪声），使得音频声音像是在水下录制的。由于问题太严重，即便经过多年的发展，这种方法还是对于提高人们在嘈杂环境中识别语音的能力束手无策。

我意识到必须采取不同的方法来解决这个问题。我们从艾伯特·布雷格曼（Albert Bregman）的理论入手。布雷格曼是位于蒙特利尔的麦吉尔大学的心理学家，他在 1990 年提出，人类听觉系统将声音组织成不同的声音流。每个声音流基本上对应一个从单一源（例如附近的朋友）发出的声音。每个声音流的音调、音量和方向都是独特的。

众多声音流（例如上述朋友在喧闹的曲棍球比赛中说话的声音）组成了布雷格曼所称的“听觉场景”。如果多种声音在同一时间处于同一频带，那么场景中最响亮的声音会压倒其他声音，这一实用原理被称为听觉掩蔽。例如，当暴雨敲打着屋顶时，人们可能不会注意房间角落处时钟发出的滴答声。这个原理被用于压缩 MP3 文件，通过消除被掩蔽的声音（例如上文提到的时钟滴答声），将文件缩小到原始大小的 1/10，而用户不会察觉。

回顾布雷格曼的工作，我们设想到是否可以构建一个过滤器，来确定一个声音流在某一时刻是否会压倒特定频带内的其他声音流。研究声音感知的心理声学家将人类的平均听力范围划分为 20 到 2 万赫兹之间的大约 24 个频带。作为分离语音和噪声的第一步，我们希望过滤器告诉我们，在某些时刻，这些频带内包含语音的声音流是否会强于噪声的声音流。

2001 年，我的实验室最先设计出这样的过滤器，可以标记出声音流是由语音还是噪声主导。利用这个过滤器，我们可根据一些区别特征，例如振幅（响度）、谐波结构（音调的特定排列）和开始时间（相对于其他声音，特定声音开始发出的

插图：Emily Cooper

相对时间），开发一套机器学习程序，进而将语音与其他声音分离。

这个原始的过滤器就是我们所说的理想二元掩码。它对每个音段中发现的噪声和语音进行标记，这些音段被称为时频单元，指在特定频带内的某个短暂时间间隔。过滤器分析嘈杂语音样本中的每个时频单元，并将每个单元标记为 1 或 0。如果“目标”声音（此处指语音）比噪声强，则记录为 1；如果目标声音弱，则记录为 0。这样，就得到了 0 和 1 的集合，分别表示样本内噪声和语音的强度。然后，过滤器去除所有标记为 0 的单元，并利用被标记为 1 的单元重建语音。为了从有噪声的语音中重建可被人们理解的句子，必须要有一定比例的时频单元被标记为 1。

我们从 2006 年开始，在俄亥俄州的美国空军研究实验室测试理想二元掩码。大约同一时间，来自纽约雪城大学的一个团队独立评估了理想二元掩码。在这些实验中，过滤器不仅可帮助听力受损的人，还可以帮助听力正常的人更好地理解受噪声干扰的语句。

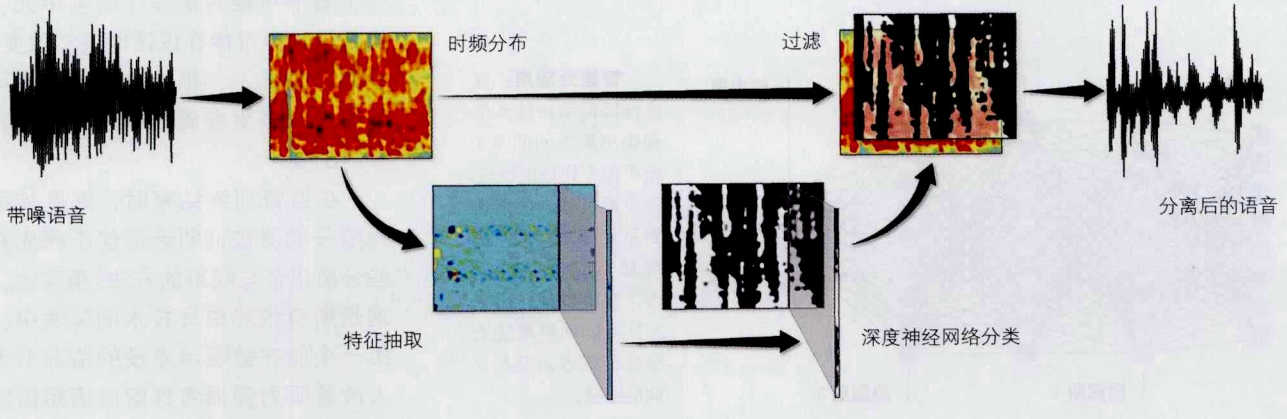
我们已基本制造出了一个在实验室内表现完美的语音过滤器。但是这个过滤器的优势条件并不切实际。根据设计，我们分别向过滤器提供了语音和噪声样本，然后再利用其对混合的样本声音进行测试。因为过滤器之前被提供了答案（这就是为什么它被称为“理想”的），过滤器知道什么时候语音比背景噪声更响亮。而在现实中，语音过滤器必须完全独立、实时地将房间中的语音和噪声分离。

尽管如此，理想二元掩码显著改善了听力受损人群和听力正常人群对语音的理解这一事实还是具有深刻的意义。它表明分类技术（一种监督学习）可以作为一种分离语音和噪声的方法，以贴近理想二元掩码。分类后，机器可以通过练习、接收反馈、吸取和总结经验等来模仿人类的学习。这与人们在小时候学习如何分辨苹果和橙子的方式是基本相同的。

在此后的几年中，我的实验室首次尝试通过分类的方法来进一步贴近理想二元掩码。大约在我们开发分类器的同时，匹兹堡卡内基梅隆大学的一个团队基于机器学习发明了自己的方法，将时频单元分类用于另一个目的：提高自动语音识别能力。后来，达拉斯得克萨斯大学的一个小组在菲利普斯·洛伊索（Philipos Loizou，已故）的带领下采用了不同的分类方法。这种方法第一次取得了意义重大的进展，使听力正常者依赖单耳特征（与通过双耳捕获声音的双耳特征相对）理解语音的能力得到提高。

但是，这些早期机器学习方法所使用的分类技术，其效果和准确性尚不足以为助听器佩戴者提供帮助。它们不能处理世界上混合在一起、复杂且不可预见的噪声和语音。为此，我们需要更强大的支撑。

在证明了我们早期分类算法的初步结果后，我们决定进行下一个逻辑步骤：完善系统，使其可以在真实世界的嘈杂环境中发挥作用，而不再针对特定的噪声和句子进行训练。这一挑战促使我们尝试一些以前从未做过的事：构建在神经网络上运行的机器学习程序，通过复杂的训练过程将语音和噪声分离。该程序将使用理想二元掩码来指导神经网络的训练。这样做是有效的。在一项包含 24 名测试对象的研究



清洁语音：为了将语音与噪声分离，机器学习程序将嘈杂的语音样本分解成被称为时频单元的元素集合。接下来，程序分析这些单元，以便提取可区别语音及其他声音的 85 个特征。然后，程序将这些特征馈送到经训练的深度神经网络，神经网络会根据类似样本的以往经验将单元分类为语音或噪声。最后，该程序利用数字过滤器，消除所有非语音单元，仅留下分离出的语音。

中，我们证明了该程序可以将听力受损人群的理解力提高大约 50%。

简单来说，神经网络是由相对简单的元素构成的软件系统，这些元素可以通过协作完成复杂的处理。（我们系统的结构大致模仿了神经元及其网络在大脑中的工作流程。）当出现新示例时，神经网络像人脑一样，可以通过调整其连接的权重进行“学习”。

神经网络的形状和尺寸多种多样，复杂程度各不相同。深度神经网络具有至少两个“隐藏”处理层，这些隐藏层不直接与系统的输入或输出相连。每个隐藏层会先改善前几层馈送给它的结果，然后基于先验知识添加新的考虑。

例如，验证客户签名的程序可能会先比较新签名与训练数据库中的样本。但是，程序也会通过训练知道，新签名与原始签名的匹配度不需要达到 100%。其他处理层可以判断出新签名是否具备原始签名的某些特定特点，例如签名倾斜的角度，或是有没有标上字母 i 上的点。

为了建立我们自己的深度神经网络，我们开始编写算法，基于每个声音振幅、频率和调制的共同变化提取可以区分语音和噪声的特性。我们确定了 85 个属性并全部采用，它们可以在一定程度上帮助我们的程序区分语音和噪声，使计算程序尽可能强大。在所有属性中，最重要的是声音的频率及其强度（高声或轻柔）。

接下来，我们训练深度神经网络使用这 85 个属性来辨别语音和噪声。训练分两个阶段：首先，我们通过无监督学习来设置程序的参数。这意味着我们把很多属性示例加载到程序中，为随后实时分类信号类型做准备。

然后我们利用嘈杂的语音样本及其经理想二元掩码处理后的相应结果，来完成第二阶段的训练，这一阶段是有监督的学习。特别是，构成理想二元掩码的 1 和 0 的集合就像一张答题纸，我们用它来测试和提高程序分离语音和噪声的能力。对于每个新样本，程序会先从带噪语音中提取一组属性。对属性（包括频率、强度等）进行分析之后，过滤器执行临时分类（是语音还是噪声），然后将结果与理想二元掩码在相同情况下确定的结果进行比较。如果结果与理想二元掩码过滤器中的 1 和 0 不同，则相应地调整神经网络的参数，以便网络在下一次尝试时得到与理想二元掩码中 1 和 0 更接近的结果。

为了进行调整，我们首先计算了神经网络的误差，测量理想二元掩码和神经网络输出层结果之间的差异。一旦计算出这个误差，我们接下来就可以用它改变神经网络连接的权重，从而在下一次进行同样的分类时减小差异。神经网络需要进行数千次这样的训练。

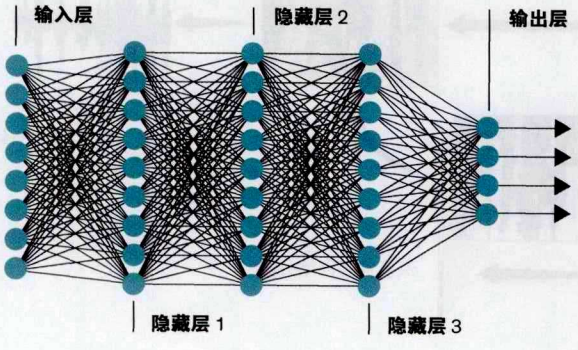
这个过程中的一项重要精化步骤是构建第二个深度神经网络。第二个神经网络接受第一个网络的馈送并微调其结果。第一个网络侧重于在每个单独的时频单元内标记属性，而第二个网络则检查特定单元附近几个单元的属性。换言之，第二个网络向第一个网络提供其正在处理的语音和噪声的语义背景，进一步提高分类的准确性。举例来说，一个音节可能跨越多个时频单元，但背景噪声可能在说话时突然改变。在这种情况下，相关的背景线索可以帮助程序更准确地分离语音和音节内的噪声。

在监督训练结束时，深度神经网络分类器被证明远远优于原先那些分离语音与噪声的方法。事实上，这是所有依赖单耳技术的算法中，第一个能在被噪声淹没的情况下大大改善听力受损者理解口语短语能

1880至1920年

1921至1953年

1984年至今



智能处理层：深度神经网络由输入层和输出层之间的两个或更多个处理层组成。信息通过输入层馈送到系统（左），输出层显示结果（右）。为了提高性能，研究人员可以调整系统的参数，或者调整层之间的连接。

插图：Erik Vrielink

1900至1945年

1952至1995年

1996年至今

微调： 在过去的150年里，技术发展改善了助听器的性能。早期的对话管（前页顶图）完全依赖于声学放大。第一个电助听器（本页顶图）使用碳膜来放大声音。采用真空管的助听器（前页中图）根据佩戴者的听力损失类型采取不同的方式处理频率。采用晶体管后，首次出现了戴在耳朵后面的助听器（本页中图）。耳蜗植入物（前页下图）彻底改变了严重听力损失的治疗方式。今天的数字助听器（本页下图）将声波转换成数字二进制代码进行处理，然后输出模拟信号供佩戴者听到。

图片（自上而下）：助听器博物馆（2）；LES LADBURY/阿拉米图片社

力的算法。

为测试在人类身上的效果，我们让12位听力受损者和12位听力正常的人通过耳机收听带噪句子的样本。样本是成对的：首先语音和噪声同时出现，然后播放经深度神经网络程序处理过的同一样本。这些包含短语的句子（如“这儿变冷了”和“他们吃了柠檬派”）混杂着两种类型的噪声，一种是稳定的嗡嗡声，另一种是许多人同时说话的嘈杂声。稳定的噪声类似冰箱运转的声音，其音频波是重复的，而且频谱的形状不会随时间变化。嘈杂背景声中加入了4男4女的说话声，以制造吵闹的鸡尾酒会效果。

实验证明，在句子经过我们的程序处理后，两组人对句子的理解程度都有很大提高。未经程序处理前，听力障碍患者只能在嘈杂的说话声中辨别出29%的词汇，但经程序加工后，他们可以听清84%的词汇。有些受试者收听原始样本时只能理解10%的词汇，而样本经加工后，这一比例提高到约90%。听力受损受试者在稳定噪声下也有类似的提高，他们的词汇理解比例从36%提升到82%。

即使听力正常的人也能更好地理解带噪句子，说明未来受益于该程序的人群可能会远远超过我们最初的预期。未经程序处理的情况下，听力正常的受试者能够理解稳定噪声下大约37%的口语词汇，经过处理后可提高到80%。在嘈杂的说话情境下，他们所理解的词汇比例从42%提高到78%。

实验最有趣的一个结果是，如果有人问：“在我们设计的程序的帮助下，听力障碍患者的听力能够比听力正常的人还好吗？”不可思议，答案是肯定的。与那些听力正常、依靠自身听觉系统分离语音和噪声的正常人相比，使用了该程序的听力障碍患者在嘈杂的说话声中所理解的词汇量高出近20%，在稳定噪声中则约高出约15%。这些结果使我们应用深度神经网络构建的程序成为目前解决鸡尾酒会问题的最好方法。

当然，程序的能力是有限制的。举例来说，在我们的样本中，遮蔽语音的噪声类型与程序进行分类训练时使用的噪声类型非常相似。为了在现实生活中发挥作用，该程序将需要快速学习如何过滤掉种类众多的噪声，包括程序未曾遇到的不同类型的噪声。例如，通风系统的嘶嘶声与冰箱压缩机的嗡嗡声就是不同的。此外，我们使用的噪声样本没有涉及房间中物体和墙壁的回声，而它们也是构成鸡尾酒会噪声问题的因素之一。

在公布了这些早期的结果之后，我们购买了一个声音效果数据库（原本用于为电影制作人设计声效），并使用其中的1万个噪声进一步训练该程序。2016年，我们发现再次接受训练的程序可以应对全新的噪声，能够切实改善听力障碍患者和听力正常者的理解力。现在，在美国国家听力及其他交流障碍研究所的资助下，我们正在推动该程序在更多的环境中运行，并让更多的听力受损者参与测试。

最后，我们相信，该程序可以在强大的计算机上接受训练，并直接嵌入到助听器中，或通过无线链路（如蓝牙）与智能手机配对，将实时处理的信号馈送到耳机。制造商定期用新噪声重新训练系统并发布新版本后，助听器佩戴者可以随之更新设备。我们已经为该技术申请了多项专利，并正在与合作伙伴（包括美国领先的助听器制造商、位于明尼苏达州伊登普雷利的斯达克听力技术公司）一起实现其商业化。

有了这个方法，鸡尾酒会问题就不会再像几年前那样令人望而生畏。我们和其他研究人员现在可以创建软件，预计能够通过更多噪声场景的训练攻克这个难题。事实上，我猜想这个过程与儿童早期学习分离语音和噪声的方式类似，即反复暴露在大量的语音和噪声环境中。有了更多的经验，这种方法一定会变得更好。这正是它的奇妙之处。它正值青年，我们还有时间。■