基于深度学习的小样本声纹识别方法

原熙文
国防科技大学 安徽 合肥 230000

【摘要】在深入研究和改进深线模型的基础上，提出了一种声带识别方法。通过改变 GMM-UBM 的阶次，确定了最经济的 m 值。利用卷积神经网络和递归神经网络进行了语音串检测实验。实验结果表明，CNN 和 ESN 能够有效地识别声带客户，识别准确率高，基于 ESN 的声带识别能够满足当前声带识别的需要。如果解决了 CNN 参数设置的问题，CNN 得到了广泛的应用。

【关键词】深度学习；声纹识别；回声状态网络；卷积神经网络

1 前言

在现代社会，随着数字时代的到来，大量的信息和数据充斥着生产生活的每一个角落。然而，随着科学技术的发展，这一安全领域也面临着诸多风险和挑战。解决身份识别问题：技术不好，容易实现，但也容易丢失、丢失和破解，这是生物技术的起源。生物特征检测是通过识别个体特征来完成的。这些独特的属性，即人格的生物属性，通常是在生物属性（指纹、静脉、面部、DNA）上进行划分的。由于其唯一性和稳定性，个体的生物学特性是认证的基础，决定了认证的安全性。与其他生物技术方法相比，语音识别具有这样的优势。你需要一个麦克风来随时随地接收语音信号，无论是电话还是电脑。正如人脸识别可以通过一个简单的算法来实现。

2 字符串的声纹识别

声纹识别是对单一语言特征的识别，即声纹识别不同于声纹识别。磁带识别是对口语的识别，可视为身份控制，而声纹识别是内容识别，如果作者有特定的身份认证，且语音认证通过，则称为“认证”。另一方面，如果我们想要识别一个未知的广告客户，并将他的声音与模板进行比较，那么他们的一致性在某种程度上是 1:1，他们的一致性是第一。

3 深度学习和培训

机械学习是最快和最活跃的领域之一，而深度学习是其真正的优势。它是人工智能的基础之一。它允许以多级抽象模型的形式表示数据。这些技术极大地改善了声纹识别、视觉目标识别、目标识别等领域的发展状况，研究了药物检测、基因组学等结构复杂的大数据集。实现语音和音频编辑。网络定期报告文本和声音等串行数据。Siri 和 Kor Tana 等系统在某种程度上是通过深度学习实现的。对数据的详细研究不是基于预先确定的方程，而是基于数据的基本参数，采用基于计算机训练的多层数据识别方法。

3.1 递归神经网络（RNA）

神经网络（RNA）是一种人工神经网络，它在一定的时间连接细胞。这允许在一段时间内动态移动。与头部神经网络不同的是，RNA 可以用其内部信息存储器处理任何输入线。

回声状态网络（ESN）是一种神经网络。隐性层的连接很少（通常为 1%），隐性神经元的连接和强度是固定的，这种网络的主要优点是，尽管其行为是非线性的，在学习过程中唯一需要改变的意义是，它在本质上不是线性的，网络能够创造一种新的时间模式。隐性神经元与突触神经元输出的关系。误差函数是参数向量的平方函数，很容易与线性系统区分开来[1]。

非参数贝叶斯公式的输出覆盖率为：在学习数据的情况下，将输出值从预测中剔除，提出了基于 ESN 核的求解方法。这种解决方案类似于（有限地）强调在一些基准中进行培训。

3.2 腹腔镜神经网络（CNN）

神经潜望镜网络（CNN）由一个或多个循环（通常是分区选择）和一个或多个完全相关的层组成，例如，作为具有多个层的标准神经网络，CNN 结构设计为使用 2D 结构进入图像。

（1）波段

卷积是 CNN 的核心。层参数由一组已经研究过的滤波器（或核心）组成。这些过滤器具有较小的允许字段，但可以扩展到输入卷的整个深度。在转发过程中，每个过滤器向上滚动输入量的宽度和高度，并通过计算过滤器的输入量与输入点的乘积来创建过滤器的二维激活图。当网络在特定的输入空间中发现某些属性时，它将学会过滤[2]。

所有有源滤波卡折叠到一定深度，形成一个完整的输出体积。因此，体积的任何元素都可以解释为神经元的输出。它在输入过程中看到一小块区域，并与同一激活图上的神经元共享参数。

（2）收集器

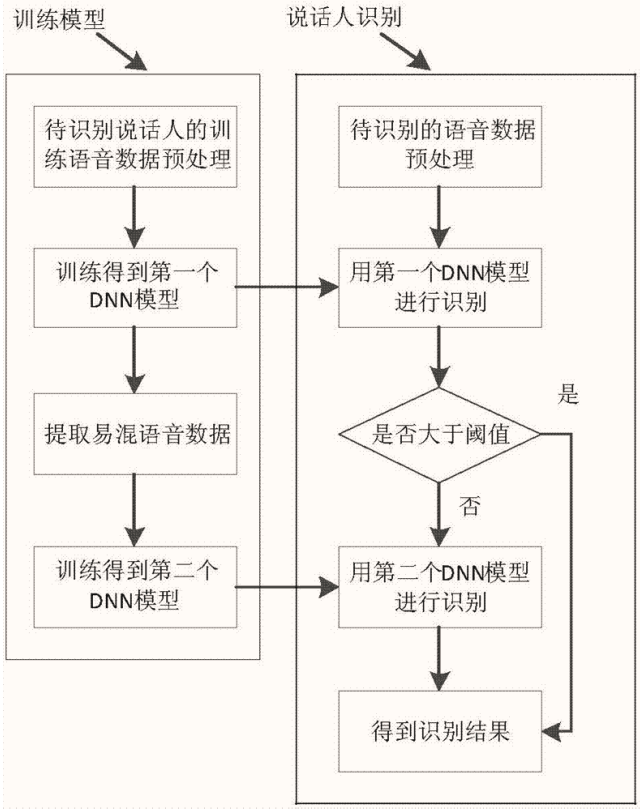
CNN 的另一个重要组成部分是下一个采样层，它由几个非线性函数来实现。最大的是将输入图像分成不重叠的矩形组，每个子组的输出最大。与其他特征相比，特征的精确排列并不重要。由于采样器在下一步可以逐步减小指定空间的大小，因此可以通过减少网络中的参数和计算量来控制仿真。通常将下一个采样层插入到 CNN 结构中。此操作提供不同的平面视图。

（3）完整连接级别

经过反复凝聚和最大样本倾倒，在完全连接层，神经元连接到上层所有活动。与一般的神经网络一样，它们的激活可以通过乘法和位移矩阵来计算。

（4）深度学习方法的改进

提出了一种新的模型，该模型将噪声抑制与Boltzmann器件相结合。根据DAE和RBM的分布情况，可以将其分为初始层DAE和初始层RBM。上层是RBM，利用RBM对原始模型进行细化和准备，提高深层神经网络的效率[3]。



4 选择深度学习的原因

进一步研究的目的是模仿人脑的思维。例如，人脑可以控制人脑的运动、感知、思维、语言等功能。通过深入学习，你可以学习如何识别声音、单词、图像和其他主题。然后电脑会做一些人们可以直接根据自己的感受来决定的事情，比如识别人脸和理解单词。我们选择深度学习来解决声纹识别的问题。主要动机是解决这个问题。表面学习造成的深度不足以模仿人脑的深层结构和抽象的人类知识。

4.1 对深度的需求

如果神经网络的深度不够，可以用定量参数来解决。然而，由于深度不够，所需参数的个数增加，使得问题更加复杂。如果深度设置为2，就足够了，但是如果每层所需的元素数量（如权重参数的数量）过大，在逻辑电路、阈值神经元等结构中，当输入模块的数量增加时，结构中所需的单元数量也会增加。通过输入数据，这些结果将呈指数增长。

深层结构可以分解。虽然很多功能不能用深层结构和小空间结构来有效表达，但大多数功能不能用深层结构和小空间结构来表达，能用深层结构表达的功能不能用平面结构来表达。

4.2 大脑结构很深

由于人脑的结构很深，对人脑的模拟算法必须有深度。人脑是一个由许多神经元组成的庞大系统。许多科学家对视觉皮层进行了广泛的研究。每个输入层都是一个特征，在更高的层次上有许多抽象的特征。这个过程是由一个训练有素的神经网络来完成的。生物神经元需要很短的时间，人脑的神经系统需要大量的神经元来处理这个过程，并不断地训练和强化，整个过程就是一个认知过程[4]。

4.3 认知过程的结构和深度模型是一致的。

人类的认知过程甚至不可能完成，但它是导致极其复杂信息的过程的一部分。这就是为什么信息处理需要一个非常深刻的框架。认知过程经常被分类。他们首先研究相对简单的概念，然后抽象出优越的特征和意义。认知过程是一个深层次的认知过程。为了用计算机模拟人类的认知过程，我们需要建立一个深层次的学习结构。

为了解决说话人识别问题，还需要建立一个根深蒂固的网络结构。本文采用Gauss-Bernoulli系统将Boltzmann与机理联系起来，在学习过程中采用Gibbs采样和对比度散射方法来更新模型参数。本文介绍了一种软件最大回归方法来细化模型参数。该模型可用作说话人识别模型[5]。

5 实际应用问题

在语音识别系统中应用高级学习模式需要考虑很多实际问题，这些问题总结如下。

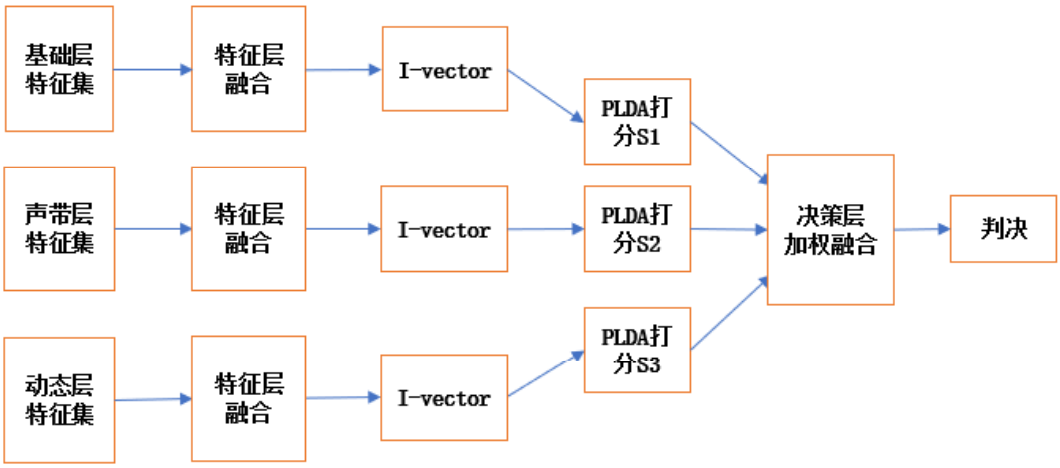
5.1 选择参数以识别扬声器

在语音识别的研究中，需要计算语音特征参数。如果选取的参数不能反映语音信号的性质，那么无论采用何种准确的模型分类和分类方法都不能提供良好的识别效果。

这些参数范围很广，如背景周期、共振峰、LPCC、线性对（LSP），MFCC不同的特征反映了不同的语音特征，在LPCC语音识别系统中，MFCC和GFCC具有较高的识别水平。本文将测试这些功能并分析它们对系统性能的影响。

5.2 了解有关数据批大小的更多信息

在深度学习模型中，每组学习数据评估模型的梯度，需要更新模型参数。一般来说，更有效的方法是对数据进行处理，这里需要指出的是，用梯度递减法不能把训练数据分解成太少的数，也就是说，尽管分类的训练数据量很大，使得梯度估计更稳定，这不会大大提高整个信念网络的学习水平，也不会使网络模型参数的更新变得微不足道。



对于等概率数较少的训练数据集，最好保证每个部分的得分等于类别数，并且每批训练数据至少有一个类别样本，从而减小了单个训练数据总梯度估计的抽样误差。对于其他训练数据，每批的大小通常较小，设置为10-100[6]。

5.3 学习速度参数

如果信任网络的深度过大，调整误差将显著增大，模型参数将不受控制。在某些情况下，这并不好，因为更新模型权重时噪声较低，但在实践中，通常很难更新模型权重。在恢复过程中，必须先降低学习系数以减小误差，然后再对模型参数进行切割以消除噪声。

6 声纹识别发展趋势

6.1 声纹识别研究朝着深度学习和端到端方向发展

语音作为语言的声音表现形式，不仅包含了语言语义信息，同时也传达了说话人语种、性别、年龄、情感、信道、嗓音、病理、生理、心理等多种丰富的副语言语音属性信息。以上这些语言语音属性识别问题从整体来看，其核心都是针对不定时长文本无关的句子层面语音信号的有监督学习问题，只是要识别的属性标注有不同。

近年来，声纹识别的研究趋势正在快速朝着深度学习和端到端方向发展，其中最典型的就是基于句子层面的做法。在网络结构设计、数据增强、损失函数设计等方面还有很多工作去做，还有很大的提升空间。

6.2 提升声纹识别系统的短时语音情况

在实际应用中，由于对基于语音的访问控制需求的不断增长，提升声纹识别系统在短时语音情况下的性能变得尤为迫切。短时语音中说话人信息不足以及注册和测试语音的文本内容不匹配，对于主流的基于统计建模的声纹识别系统是一个严峻的挑战。

6.3 改进现有的深度说话人学习方法

目前采用的深度说话人识别方法首先利用神经网络提取前端的帧级特征，然后通过池化映射获得可以表示说话人特性的段级向量，最后采用 LDA/PLDA 等后端建模方法进行度量计算。

相对于传统的 i-vector 生成过程，基于深度学习的说话人识别方法优势主要体现在区分性训练和利用多层网络结构对局部多帧声学特征的有效表示上。如何进一步改进现有的深度说话人学习方法是现阶段的一个研究热点。

6.4 深度对抗学习在声纹识别技术中的应用

生成式对抗网络 (GAN) 的主要目的是用在数据生成、降噪、等很多场景里面。它还被用在领域自适应里面，形成一个新的分布。第三个广泛的应用是生成对抗样本，这会对分类系统产生大的困扰。很多研究者用对抗样本攻击机器学习的系统，在原始数据上增加一些扰动，生成样本，经过神经网络之后就有可能识别成完全不同的结果。这个思想在图像处理领域非常活跃，会造成错误识别，引起了自动驾驶，安全等领域的研究人员的广泛关注。

在语音领域,GAN 可以用在语音识别、口音自适应上，通过多任务学习和梯度反转层来进行口音或信道的自适应，然后加上其他方法可以得到较好的效果。声纹识别也存在各种不匹配的问题，在声纹识别上也可以使用这一思想。同样的思想也用在了 TTS 语音合成领域，目的是把不同的音素解耦成说话人，风格等，去除噪声对建模的影响。

6.5 深度嵌入学习是进行声纹识别和反欺骗的一个重要途径

说话人识别和欺骗检测近年来受到学术界和业界的广泛关注，人们希望在实际应用中设计出高性能的系统。基于深度学习的方法在该领域得到了广泛的应用，在说话人识别和反欺骗方面取得了新的里程碑。然而，在真实复杂的场景下，面对短语音、噪声的破坏、信道失配、大规模等困难，开发一个鲁棒的系统仍然是非常困难的。深度嵌入学习是进行说话人识别和反欺骗的一个重要途径，在这方面已有一些著名的研究成果。如之前的 d-vector 特征和当前普遍使用的 x-vector 特征。

7 结束语

通过以上分析可以得出以下结论：GMM-UBM 的阶数对语音识别有积极的影响，但问题是，阶数越高，GMM-UBM 算法越复杂，计算难度越大。使用古典声乐学习方法时，最好使用 64 倍的顺序。通过比较不难发现，传统的基于高斯混合的检测方法的检测精度比 CNN 和 ESN 差，CNN 的检测精度比 ESN 高。卷积神经网络是可行的，CNN

在语音识别系统中的性能较好，由于神经网络处理和参数设置的困难，需要进行大量的实验，恢复以提高磁带检测的准确性。

【参考文献】

[1]李靓，孙存威，谢凯，等．基于深度学习的小样本声纹识别方法[J]．计算机工程，2019,v.45;No.498(03):268-273+278.

[2]韩侣，周林华，马文联，等．基于深度学习的小样本声纹识别研究[J]．应用数学进展，2020(1):30-37.

[3]张颖，徐志京．基于深度学习的帕金森患者声纹识别[J]．计算机工程与设计，2019,040(007):2039-2045.

[4]杨楠．基于深度学习的说话人识别研究与实现[D]．郑州大学，2019.

[5]郑博文．基于深度学习的生物特征识别方法研究[D]．2019.

[6]池春婷．基于UBM和深度学习的说话人识别方法研究[D]．大连理工大学，2020.