中国科学: 信息科学 2013 年 第 43 卷 第 12 期: 1578–1597
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

评 述 中国科学院学部 科学与技术前沿论坛 信号与信息处理专刊

互联网时代语音识别基本问题

柯登峰*, 徐波

中国科学院自动化研究所, 北京 100190
* 通信作者. E-mail: dengfeng.ke@ia.ac.cn

收稿日期: 2013–06–20; 接受日期: 2013–11–13
国家重点基础研究发展计划 (批准号: 2013CB329302) 资助项目

摘要 语音识别技术经过半个世纪的积累, 于近年来达到大规模商用水平. 本文概括了统计语音识别理论的发展状况, 并单独介绍了深度神经网络在声学建模、语言建模、多语言共享、语义识别等方面的卓越性能. 深度神经网络的性能优势引起了我们强烈的兴趣. 通过回顾类人听觉信息处理对深度神经网络的改进作用, 我们意识到, 深度神经网络与类人听觉信息处理相结合, 必将推进语音识别技术的进一步发展. 反过来, 深度神经网络技术在语音识别中的进步, 也必将推动类人听觉信息处理技术的进步. 语音识别技术后续发展的重点是对深度神经网络的结构和训练算法的改进使之更好地实现类人听觉. 最后, 我们分析了采用深度神经网络模拟人类听觉的抗噪修复机理和听觉关注机理的可能性.

关键词 信号处理 语音识别 神经网络 深度神经网络 类人听觉

1 引言

语音识别技术经过半个世纪的酝酿和积累, 在最近几年达到了大规模商用水平. 2009 年, 国际著名的信息技术咨询公司 Gartner 公司首次将语音识别技术列入复苏 (slope of enlightenment) 阶段的尾声, 并指明语音识别技术将在 5~10 年内获得大规模商用 [1]. 2011 年, Gartner 公司进一步修改新兴技术炒作周期 (emerging technologies hype cycle) 曲线, 将语音识别的大规模商用时间提前到 2~5 年内 [2]. 这一更改预示着语音识别技术已经近乎成熟. 与此同时, 信息技术领域相关行业的巨头企业纷纷瞄准了语音识别技术竞相出手. 也是在这个时期, 语音识别、语音翻译、语音问答、语音控制等一系列概念伴随着语音产品走进千家万户. 语音识别技术重新成为信息产业的热点问题.

在众多的行业应用中, 移动互联网首先成为连续语音识别技术大规模商业化最重要的应用环境. 从 2010 年起, 谷歌带头在安卓系统上开放了在线语音识别引擎接口, 用户可以利用此接口进行音字转换, 或用于控制智能手机的各种操作. 短短两年时间内, 科大讯飞、Apple、百度、盛大、云之声、腾讯、中科院自动化所等公司或单位相继推出了在线语音识别引擎和应用软件. 今年来, 科大讯飞、云之声、中科院自动化所又相继推出了适合手机平台使用的离线语音识别引擎. 短短数年, 互联网搜索巨头、网游巨头、通讯巨头和学术带头单位纷纷进军语音识别的产业化应用, 掀起语音识别的应用热潮.

在产业化进程热火朝天地进行的时候, 语音识别核心技术的研究也在快速推进. 声学特征提取技术和声学建模技术经过几十年的发展, 形成了很多有效的经验和技术方法. 语言知识的表示体系和解

引用格式: 柯登峰, 徐波. 互联网时代语音识别基本问题. 中国科学: 信息科学, 2013, 43: 1578–1597, doi: 10.1360/N112013-00078

码搜索技术也获得长足的发展. 语言模型与声学模型的耦合程度越来越高, 搜索策略与搜索空间的描述方式随着语言知识的表达体系和软硬件体系结构迅速变更. 特征提取、声学建模、语言建模、搜索策略从独立优化走向联合优化. 深度神经网络 (deep neural network, DNN) 的出现带来了语音识别的新一轮突破, 引发了新一轮的思考. 深度神经网络在类人听觉信息处理方面取得了许多成就. 深度神经网络与类人听觉信息处理融合, 必将成为语音识别的重要研究方向.

2 统计语音识别理论和方法

2.1 声学特征和建模技术

最近 40 年, 语音识别的特征提取技术和声学建模技术取得翻天覆地的变化. 声学特征抽取和声学建模技术从相对独立逐步走向融合统一, 声学模型逐步拥有了对原始特征进行二次优化抽取的能力. 音高感知、响度感知、声道长度、时间掩蔽、频率掩蔽等一系列的知识逐步加入特征提取算法之中, 并与声学模型紧密结合在一起.

声学特征提取和声学建模技术曾经是两种相对独立的技术. 声学特征提取旨在从复杂多变的语音波形中抽取相对稳定的最能代表发音内容的特征向量. 而声学建模意在统一描述声学特征的内在规律.

许多研究表明, 声学特征的优化技术和声学建模的优化技术在很多情况下是可以互相转换的. 对声学特征进行线性的声道长度归一化 (VTLN) 等效于对声学模型进行约束的最大似然线性回归 (CMLLR)[3]. 对声学模型的每一种区分度训练算法, 均可以找到对应的声学特征变换方法. 典型的如: MPE 和 fMPE[4], BMMI 和 fBMMI[5].

近几年来, 深度神经网络使得特征提取和声学建模走向深度融合. 它以多层的底层网络对原始的声学特征进行层层变换, 形成更好的声学特征, 并在最后的顶层网络中进行区分度训练, 通过将特征优化抽取和声学建模融合在同一个网络中进行优化, 获得了极大的性能提升.

2.1.1 声学特征提取

模拟人耳听觉是声学特征提取技术成功的重要原因. 梅尔倒谱 (Mel-scale frequency cepstral coefficients, MFCC) 由于考虑了听觉音高的距离, 获得了比线性预测倒谱 (linear prediction cepstrum coefficient, LPCC) 更好的性能. 早期的研究表明, 人耳对音高的感知并不是在频率域上等距离的. 在 1KHz 以下的频段, 人耳对音高感知接近线性等距离; 而在 1KHz 以上的频段, 人耳对音高的感知接近对数等距离. 根据 Paul 的记载 [6], Bridle 和 Brown 等人最早利用人耳的这种特性创造了 MFCC 特征 [7], 并在孤立词语音识别中获得成功. 2000 年, 欧洲电信标准协会 (ETSI) 将 MFCC 提取流程标准化, 并提供了 16, 11, 8 K 条件下的 3 套规格标准 [8]. 听感线性预测特征 (perceptual linear predictive, PLP) 也是因为考虑了人耳听觉特性而获得成功. 1989 年, Hermansky 在线性预测特征中加入了可以模拟人耳听觉掩蔽现象的 Bark 滤波器, 以及可以模拟人耳听觉响度的等响曲线, 获得了性能大幅度提升, 并将这种特征命名为听感线性预测 (PLP) 系数 [9]. 后来, Woodland 等人发现把 PLP 中的 Bark 滤波器组改为与 MFCC 一致的 Mel 滤波器组, 可以在多数情况下获得更优的性能 [10], 形成了后人所熟知的 MF-PLP 特征.

特征的处理技术同样对性能产生重大影响. 语音识别技术要面向应用, 必然会遇到各种噪声, 因此需要各种各样的抗噪技术. 一类抗噪技术是寻找噪声条件下的稳定因素, 例如 Hermansky 提出的

RASTA 技术 [11]. 它的理论基础是信道噪声虽然改变了频谱, 但并不能改变频谱的局部变化趋势. 通过提取语音的相对谱, 并将它应用到 PLP 特征提取中, 可以极大地减轻信道带来的识别错误, 但在纯净录音条件下性能略有下降. 另一类抗噪技术是将受到噪声污染的录音尽量恢复成纯净状态, 典型的如 MMSE[12] 和 VTS[13] 技术. 近年来新出现的 PNCC[14], PNPLP[15] 等也是类似的思想. 它们通过对背景噪声细化估计, 对听觉的时间掩蔽和频率掩蔽效应进行深度挖掘, 从而获得比以往更好的抗噪效果.

VTLN 技术可以很好地消除说话人之间发声器官差异带来的影响. 研究表明, 在梅尔域进行的声道长度归一化具有最优的识别性能 [16]. 对声道长度的快速准确估计直接关系到该技术能否被实际系统所采用. 通常而言, 直接采用似然概率最大化无法获得声道长度的解析解, 但是通过增加约束条件, 可以获得快速高效的近似解 [17]. Hermann Ney 等人证明了声道长度归一化可以等效于在倒谱域的线性变换, 进而推导出在倒谱域变换的解析解 [18]. 在最大似然目标函数下, 对声学特征进行线性的 VTLN 和对声学模型进行 CMLLR 是等效的 [19], 因此线性 VTLN 和 CMLLR 带来的性能提升不能叠加 [20].

判别特征提取技术是非常重要的技术. VTLN 技术本质上只能解决人与人之间声道长度引起的共振峰位置的差异. 人和人的其他差异并不能通过此技术消除. 另一种解决消除说话人之间的差异的方法是直接从特征中抽取说话内容的特征, 忽略其他无关的特征. 最初, Hermann Ney 等人将线性判别分析 (linear discriminant analysis, LDA) 引入到语音识别特征的提取 [21]. 该方法通过一个线性变换将原始特征投影到另一个空间上, 使得类间方差和类内方差的比值最大化, 从而使得新特征对所给定的类别最具有区分性, 人和人之间的个性化差异, 以及信道不同带来的差异都可以被最小化. 由于 LDA 引入了等方差假设, 这种假设在某些情况下并不能使得类别区分性最优, 甚至可能导致更加糟糕 [22], 因此有部分人研究在异方差条件下的线性判别分析 (HLDA) , 然而并不是每一种 HLDA 技术都是有效的, 有些时候 HLDA 效果甚至还不如 LDA. Kumar 等人提出的 HLDA 算法将 LDA 的思想推广到异方差条件下, 并在隐马尔科夫模型 (hidden Markov model, HMM) 框架下推导出能够使得似然概率最大化的线性判别变换 [23] 的计算方法并取得成功. 上述判别特征提取技术是在孤立的帧上获得区分性的最大化. 由于语音识别所用的 HMM 模型假设了帧间的独立性, 这一假设与实际情况不符, 因此将多帧串在一起进行时序化的区分度训练可以进一步提升系统的性能. 这类时序化区分度训练算法通常包括 bMMI 和 MPE 等, 其对应的特征层面上的区分度训练算法为 fBMMI 和 fMPE[4,5].

神经网络 (neural networks, NN) 具有更复杂的非线性处理能力. 它在声学特征提取方面性能优越. 采用神经网络进行语音识别并不是近年来的首创, 早在上世纪末期, 神经网络便已用于语音识别, 主要包括两种方法: Tandem 方法和 Hybrid 方法. 早期的研究表明, 利用神经网络的输出作为特征可以使得同一类别的特征输出更加集中, 从而更加适合采用高斯混合模型 (Gaussian mixture model, GMM) 建模, 这种利用神经网络输出作为特征训练 GMM-HMM 模型的方法称为 Tandem 方法. 实验表明这种方法可以使识别性能大幅度提升 [24~26]. 与 Tandem 方法不同的是, Hybrid 方法直接采用神经网络模型替换传统 GMM-HMM 模型中的 GMM, 这类方法也可以使得识别率获得大幅度提升. 语音是一种时序变化的信息. Hermansky 等人最早提出一种将 MLP 和长时间变化轨迹相结合的方法, 并称之为 TRAP 特征 [27]. 后来利用多帧拼接获得长时间特征方法被证实具有显著效果.

2.1.2 声学模型训练

隐马尔科夫模型 (HMM) 在 20 世纪 60 年代末 70 年代初被引入语音识别领域, 并因其简单有效的特性而受到语音识别领域专家们所喜爱. 以高斯混合模型 (GMM) 为隐藏状态的声学模型在最近二

十年取得了突飞猛进的发展.

Baum-Welch 算法、extended Baum-Welch 算法是语音识别领域最为基础的模型参数估计算法, 通常用于估计最原始的声学模型. 针对说话人之间的频谱差异, 人们研究了一系列说话人自适应算法, 包括矢量量化法 (vector quantization)、层次化谱聚类法 (hierarchical spectral clustering)、概率谱映射法 (probabilistic spectral mapping)、贝叶斯自适应法 (Bayesian adaptation) 等一系列方法 [28]. 这些算法多数不适合用于连续密度 HMM 模型, 只有贝叶斯自适应算法与连续密度 HMM 模型可以轻松融合, 并由此变化形成了后人所熟知的最大后验概率法 (MAP) [29]. 该算法所产生的效果是, 随着特定人数据的增多, 原始的模型将被逐渐修改成特定人的模型, 从而获得说话人自适应的效果. 和 MAP 不同, Woodland 等人提出的最大似然线性回归法 (MLLR) 可以采用很少的数据 (10 秒以上) 估计模型的均值变换矩阵, 并获得很好的性能提升 [30]. Gales 等人将此技术丰富成最大似然线性变换 (MLLT) 技术 [31].

产生式 (generative) 方法和区分度 (discriminative) 方法是统计模型的两种最主要的训练方法. 产生式方法通过学习观察向量的联合概率分布, 并利用贝叶斯准则转换成后验概率从而使得模型具有识别能力. 区分度训练算法则直接优化类别的后验概率, 无需经过贝叶斯转换这一层中间步骤, 因而可以获得比产生式方法更有效的区分能力. 这类具有最佳区分能力的识别器通常包括支持向量机、条件随机场、最大熵马尔科夫模型等. 对语音而言, 简单的区分度训练方法尚不够用. 语音是一种时序化的序列, 语音识别问题是时序化的一个个模型的识别问题, 因此一个模型的识别错误将会引起后续识别的连锁反应错误. 另外, 由于语音发音单元通常被假定为一个个的 HMM 模型, 这就人为地为语音增加了帧间的独立性假设. 这种假设削弱了模型的识别能力. 因此, 语音识别的区分度训练通常是在句子层面上进行优化的时序化的区分度训练算法, 主要包括三大类: 最大互信息 (MMI) 方法 [5,32]、最小分类错误率 (MCE) 方法以及最小音子/单词错误率 (MPE/MWE) 方法 [33]. MMI 和 MCE 均是在句子级大粒度单元上的区分度训练方法, 而 MPE/MWE 则是在音子/单词等小粒度单元上获得的区分度训练方法. MPE/MWE 方法以音子/单词的粗精度 (raw accuracy) 加权特征观察到标签的后验概率. MMI 和 MPE/MWE 均可看成是在模型上使得后验概率损失最小化, 而 MCE 则是从分类损失出发使得损失最小化. MMI 和 MCE 均是在句子层级的大粒度上优化目标, 而 MPE/MWE 则是在音子/单词的小粒度上优化目标. MMI 优化的目标是乘积的形式, 而 MCE 和 MPE/MWE 的目标则是求和的形式. 虽然三者表现迥异, 但在特定约束条件下却可以整理成统一的形式. XiaoDong He 等人 [34] 将三者进行若干等价变换, 最终统一成相同的形式, 从而使得它们均可以采用基于 GT (growth tansformation) 的参数优化算法优化, 从而使得训练过程更加的快速和稳定. 这三种区分度训练均有多种不同的实现方法. 采用不同的方法实现区分度训练, 获得的性能提升有很大不同 [5,32]. 通常区分度训练是在词格上进行的, 三种方法孰优孰劣, 也很难有定论. IBM 的研究人员研究似乎表明 [35], 采用粒度更小的 sMBR 可以获得最大的性能提升.

2.2 统计语言知识表示和搜索技术

2.2.1 语言知识表示

早在 1959 年, 第一批语音识别专家 Fry 等人便已提出, 机器语音识别需要一个声学系统和一个语言系统 [36]. 声学系统负责将声音转成音素, 而语言系统则运用语言知识层层纠正声学系统的识别错误. 当时的学者把语言系统分成三个层次. 第一个层次是音素 (phome) 之间的连接关系与连接概率. 它帮助我们纠正一些不可能存在的发音. 第二个层次是词素 (morpheme) 之间的连接关系与连接概率.

它帮助我们纠正一些不可能的词语形态构造. 第三个层次是词 (word) 之间的连接关系和连接概率. 它帮助我们纠正一些不可能的词语连接, 从同音词中选出最佳的词语. 在现代连续语音识别系统中, 词素之间、词之间的统计概率模型通常被称为语言模型, 而音素模型不被采用, 改而使用发音词典约束每个词素或单词的发音.

语言模型是最重要的语言知识. 它以统计概率模型为主导. 多元文法 (N-gram) 模型在语音识别领域应用最广. 常见的有二元文法、三元文法、四元文法和五元文法. 一般来说, 从二元文法到三元文法, 大约可以获得 50%的性能提升, 从三元文法到四元文法, 可以进一步获得 5%的性能提升, 而四元文法到五元文法几乎没有提升. 多元文法的优点是训练方法简单、识别性能高效. 它的缺点是无法描述长距离约束, 需要大量数据才能获得充分训练. 因此, 有人提出基于聚类的语言模型, 也有人提出采用神经网络构建语言模型 [37,38]. 这两类模型均可以在较少数据情况下获得优于多元文法的语言模型. 基于聚类的语言模型通过将类似的词汇聚成同一类别, 从而使得少量数据条件下可以获得较为充分的训练. 与词语聚类相类似, 神经网络也可以很好地学习词语之间的语义关系 [39]. 早期的神经网络语言模型通常以离散的二进制码表示输入词的编号. 这种编码方式使得输入维数随着词表大小线性提升. 对数十万的大词典来说是个巨大灾难. 词哈希 (word hash) 通过将单词拆开成字母的组合, 可以在很少的冲突条件下采用低维字母组合的向量描述巨大的词表 [39]; 而实数向量表示法 (real-vector) 则把词表中离散的词语映射成连续的实数向量表示, 距离相近的向量具有相近的语义 [38]. 神经网络的训练速度是构建神经网络语言模型的瓶颈. 为了提高训练速度, 文献 [37] 通过强制把词表映射成规定类别数目的词类获得训练速度和困惑度性能的提升, 实验还表明, 通过将传统的多元文法模型和神经网络语言模型插值的方法, 可以使得困惑度进一步获得改善. 神经网络语言模型已经被证实具有比多元文法模型更优的性能. 但由于其复杂性太高, 无法直接融合到实时语音识别系统中, 通常用来对 N-best 结果进行重新打分 [40].

多元文法模型也有各种各样的优化算法. 它可以进行自适应, 以适应特殊的领域应用和特殊的人群 [41]. 它也可以结合声学模型和解码结果进行区分度训练, 以获得更佳识别性能 [42]. 为了适应超大规模语言模型训练的需要, 人们提出了各种大规模语言模型的训练算法 [43,44].

语言知识从原来的前缀树 (TRIE) 表示, 逐步转变成融合了语言模型、发音词典、上下文约束、声学建模单元多层知识的加权有限状态机 (WFST) 网络. 通过这种语言模型和声学模型间的强耦合, 搜索空间得以最大地消除冗余, 语言概率得以打散分配到各个建模单元上, 从而获得识别速度和识别性能的提升. 语言模型从独立于解码算法到结合解码算法的向前看技术、推送技术, 从无约简到高度约简, 从静态构建搜索空间到动态构建搜索空间, 产生了一系列变化.

语言模型从简单的统计到具有类别信息的统计, 从短距离的约束到长距离的约束, 从话题无关到话题相关的自适应, 从单独的困惑度优化到错误率最小化的区分度优化, 从独立于声学模型的优化到与声学模型高度融合, 从小规模模型训练到超大规模模型训练, 产生了许多改进算法.

2.2.2 搜索解码算法

语音识别的解码算法可以分为一遍搜索、多遍搜索. 而一遍搜索技术又分为基于词树的搜索、基于动态 WFST 的搜索、基于静态 WFST 的搜索. 基于词树的搜索技术消耗内存最小, 搜索速度最慢, 解码算法相对复杂, 通常需要考虑模型内扩展、模型间扩展、词间扩展等不同扩展方式, 需要各种预测技术和多种裁剪门限帮助快速剪枝. 基于静态 WFST 的搜索技术消耗内存最大, 搜索速度最快, 由于所有扩展都统一成有限状态机上的弧, 解码器设计相对简单, 有利于集中精力进行优化. 由于内存消耗巨大, 也有人提出动态 WFST 的搜索技术, 其要点是让 WFST 的组合 (compose) 操作在解码时进

行, 根据需要扩展, 从而不需要巨大的内存空间保存静态 WFST 网络. 语音识别的搜索空间通常由文法 (G)、发音 (L)、上下文 (C)、状态 (H) 四个层次构成, 在不同层次实现动态 WFST 构建, 需要的内存和解码技术都相应不同. 也有研究人员研究将三元文法 (G3) 一分为二的方法, 通过一个一元的文法 (G1) , 可以构造很小的解码空间 H ∘ C ∘ L ∘ G1, 在解码的时候, 采用 G3/G1 修正解码的语言模型得分 [45].

为了加快语音解码速度, 快速高斯计算、集束搜索等策略依旧可以使用, 随着 SSE 指令集从 SSE2 到 SSE4 的升级, CPU 的并行处理能力越来越强. 通过 SSE4 指令, 识别器可以在不损失识别性能的条件下完成 16 路并行计算. 利用 GPU 的超强计算能力, 可以获得更大幅度的速度提升.

多个解码器的识别结果融合, 可以获得更好的识别效果, 通常可以采用 ROVER 技术 [46] 或 SCARF 技术 [47].

近年来, 将语言模型、发音词典、声学建模单元统一编译成加权有限状态机网络 (WFST) 进行解码成为主流, 解码搜索算法也相应得以改变. 从慢速系统到实时系统; 从单纯的软件优化到软硬件优化; 从 CPU 优化到 GPU 优化; 从帧同步计算到批同步计算; 从单一系统到多系统融合. 统计语言知识表示和搜索解码技术经历了许多变化, 产生了许多有用的方法.

3 深度神经网络模型

上述介绍的以 GMM 为隐藏状态的 HMM 模型在语音识别领域盛行了数十年 [48], 随着特征提取技术的成熟和区分度训练算法的出现, 最终达到了可以大规模商用的水平. 而最近 5 年来, 这种高度商业化的技术又逐步被深度神经网络所超越. 深度信任网络是最早取得成功的深度神经网络. 它由多层的特征提取层和一层特征分类层 (softmax layer) 构造而成. 在底层的多层特征提取层中, 神经网络模拟神经突触的生物电的二进制传递方式, 通过层层传递信息, 原始的特征在传递过程中越来越有组织规律, 越来越有利于模式识别. 在最顶的一层网络中, 通过带有标签的区分度训练, 使得最终的信息被分成若干人为类别. 它既是一个特征提取和优化的网络, 又是一个特征分类和识别的网络. 深度信任网络之后又出现了许多十分有效的新型深度神经网络. 鉴于深度神经网络的优异性能, 以及对语音识别带来的巨大影响, 我们在此单列一章对它进行全面介绍.

3.1 语音识别面临的难题

语音识别技术在声学特征提取、声学建模、语言知识表示、搜索解码算法等方面都取得大量突破. 这些突破使得语音识别技术大规模应用成为可能, 但这并不意味着它已经解决了所有难题. 相反, 历史遗留下来的大量难题至今尚未解决.

美国标准技术研究所 (NIST) 组织的语音识别评测是国际上最有名的语音识别评测. 1987 年, 德州仪器公司 (TI) 向 NIST 提供了其录制的资源管理 (resource management) 语音库, 该库采用 991 个词结合模版文法生成了大约 2800 句话, 由录音人朗读完成. 在接下来的 4 年中, 这种按照文法朗读录音的词错误率从 20.7%下降到 3.6%. 从 1991 年开始, NIST 组织挑战航空旅游信息系统 (ATIS) 语音库. 该库是面向航班自动查询和应答服务的即兴录音, 表达方式灵活, 词汇进一步增大到数千词. 经过 4 年时间, 词错误率从 15.7%下降到 2.5%. 从 1992 年开始, NIST 组织挑战大词汇量统计语言模型任务. 该项任务包括了 5K 词汇量和 20K 词汇量两项任务, 录音的主要风格依然是朗读式, 录音文本来源于华尔街日报 (WSJ) , 后来又追加了北美商务 (NAB) 的内容. 由于词汇量增大, 当年的最佳成绩分

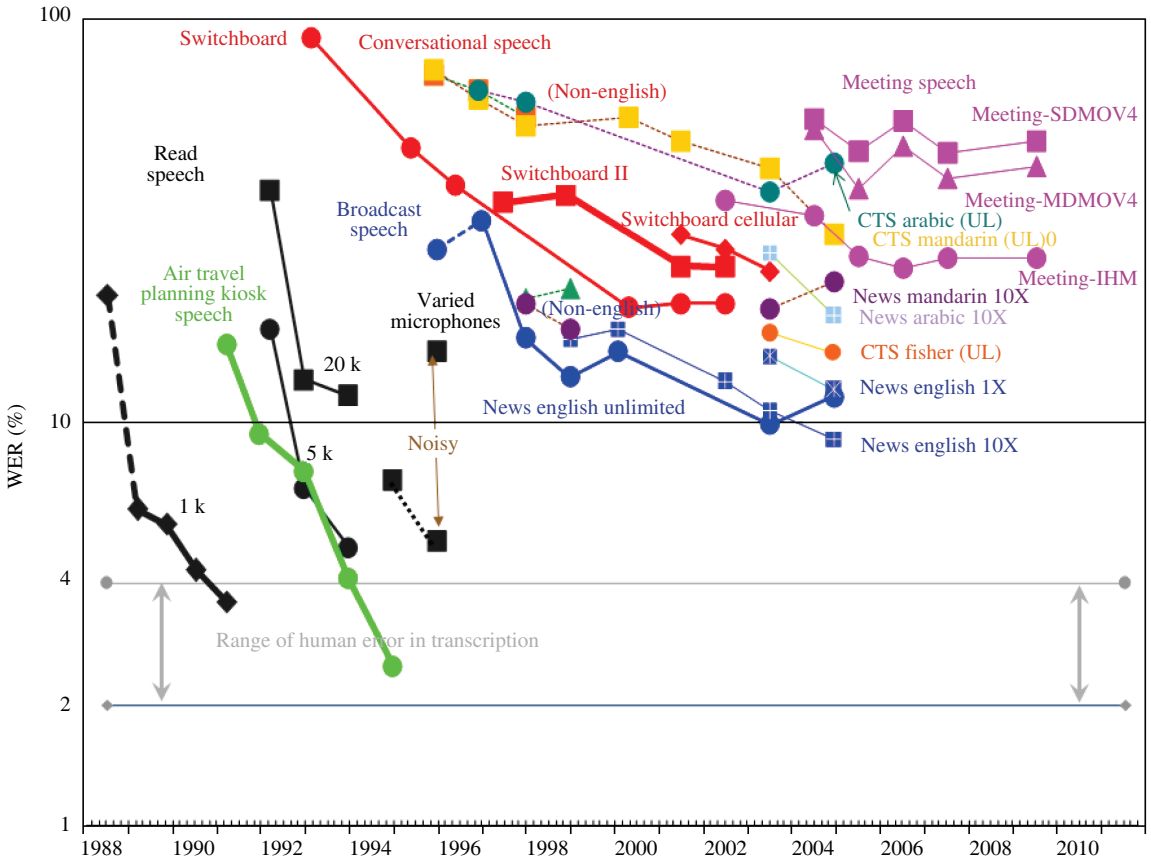


图 1　NIST 语音识别基准测试历年性能一览

Figure 1　NIST STT benchmark test history

别是 17.1% (5K 测试集) 和 32.9% (20K 测试集) 的错误率. 1995 年, 当时 5K 测试集的错误率已经下降到 5.1%. NIST 又组织评测了远讲录音和近讲录音的性能差异, 可以看到录音条件从近讲变到远讲时, 错误率从 5.1%急剧上升到 15.1% (见图 1 varied microphones) . 同一年, NIST 又组织了新闻广播语音的挑战, 数据来源于无线广播和电视广播, 从朗读式到即兴式, 从流利到不流利, 从安静环境到噪声环境, 测试条件复杂多变, 给语音识别带来极大困难. 直到 2003 年, 新闻广播录音的错误率终于突破到 10%的界限以下. 另一个更具挑战性目标是交换机 (switchboard) 语音库. 该库采集了在电话信道上就特定话题进行讨论的录音, 其初衷是为了进行话题检测和说话人检测研究. 1993 年, 人们第一次在此语音库上进行语音识别, 错误率高达 90%. 1995 年, 人们将错误率改善到 48%. 到 2001 年, 错误率又进一步下降到 19%. 由于长期以来新闻广播识别错误率远远超过人的标注错误率, NIST 开始着手考虑选择一种复杂环境的特定场景作为突破, 于是在 2002 年选定了会议记录领域. 会议录音的特点是存在多个人重叠交叉说话, 话题多样, 说话内容随意, 这些现象导致多年来会议录音的识别错误率一直居高不下, 多数条件下错误率超过 30%甚至超过 50%.

图 1 概括描述了 20 多年来 NIST 组织的各种评测的最佳性能变化情况. 虽然书面语或准书面语的连续语音识别精度已经超过了 90%, 但是, 语音识别依然面临着若干重大问题没有解决. 这些问题主要包括:

1) 环境噪声问题. 当信噪比下降时, 识别错误率成倍增长. 简单地把录音条件从近讲录音变成远

讲录音, 识别错误率发生了 3 倍增长. 而当前的各种抗噪算法, 虽然在一定程度上改善了噪音条件下的识别性能, 但是却使得高信噪比条件下错误率有所增加. 这两者是一组需要慎重权衡的矛盾.

2) 口语语音问题. 当说话人离开预先设定的文本时, 难免产生不流利现象和不符合语法规范现象, 发音通常没有朗读语音清晰饱满. 这些现象导致了声学模型和语言模型的双重不匹配, 从而导致识别率急剧下降.

3) 多说话人问题. 当多个说话人同时说话时, 目前尚没有一种好的算法可以将重叠部分分离开, 也没有一种支持多目标同时解码的语音识别算法.

深度神经网络的使得语音识别性能提升到一个全新的水平. 这种新型模型在多个领域应用均获得成功的效果. 它的出现似乎使上述问题看到了新的希望.

3.2 深度神经网络的提出

深度神经网络 (DNN) 源于专家乘积 (PoE) 系统, 与传统的专家求和 (SoE) 系统有本质差异. GMM 的概率输出由各个成分的概率求和而成, 导致它无法构造比正态分布更加陡峭的分布, 从而削弱了类别之间的区分能力. 专家乘积系统的决策结果由多个成分的概率相乘而成, 使得它在构造陡峭的分类边界方面有特殊的优势.

按照 Hinton[49] 的说法, 深度神经网络指的是在输入层和输出层中间, 含有多于一个隐含层的前馈型的神经网络. 深度信任网络 (deep belief network, DBN) 是深度神经网络中最早在大规模语音识别应用中取得成功的一种.

2002 年, Hinton[50] 在解决专家乘积系统问题时提出了一种快速使得对比散度 (contrastive divergence, CD) 最小化的方法. 这种快速算法成了训练高密度连接的约束波尔兹曼机 (RBM) 的基础. 2006 年, Hinton[51] 应用该算法训练多个波尔兹曼机, 通过层层叠加构成深度信任网络 (DBN) , 最后通过自顶向下的细调 (fine-tuning) 获得优化的 DBN 网络. 这种网络在图像识别中获得了比当时区分度训练更好的性能. 从此以后, 深度神经网络的研究逐步流行起来.

深度神经网络的训练通常包括预训练 (pre-training) 和细调 (fine-tuning) 两个步骤. RBM 预训练是在特定准则下进行的非监督训练. 这种预训练的巧妙之处在于构造一种具有特殊能力的双向的神经网络. 这种网络通过可见状态接受输入, 接受的输入经过非线性变换后激活隐藏状态, 再由隐藏状态逆向生成原始输入. 如此一个来回, 相当于一次编码过程再加一次解码过程. 求解这个编码方案和解码方案的方法可以有多种. 最直观的想法是要求编码再解码的结果必须与原始输入相一致. 然而光有这个条件是不够的, 神经网络的可见层和隐藏层之间的关系并不明确, 因此需要附加约束条件. 一种方法是从系统得分的角度出发, 以系统的似然概率最大化为目标进行训练 [50]; 一种方法是从系统的误差出发, 以系统的对比散度最小化为目标进行训练 [50]; 还有一种方法是从系统的信息丢失出发, 以输入层和输出层的互信息最大化为目标进行训练 [52]. 当一个层次训练完成时, 只保留编码方向的弧, 删除解码方向的弧, 然后用同样的方法训练后继层次的网络, 最终形成深度神经网络. 采用不同的准则进行预训练将获得不同的网络参数, 不同的网络参数极大地影响最终的识别性能. 通过一层一层叠加预训练, 原始特征被组织得越来越有结构化信息. Erhan 等人 [53] 通过图像的方法形象地描述了这种现象. 在他的实验中, 10 个阿拉伯数字被送入 DBN 网络进行预训练, 并采用激活最大化 (activation maximization) 技术进行画图, 可以看到, 处于最底层的网络学习到的是简单的横线、竖线、斜线和点等基本笔画信息, 处在第二层的网络则学到了简单的笔画组合, 而在第三层的网络则几乎学会了所有复杂的组合结构. 这些复杂结构已经包含了 10 个数字的各种写法. 最后, 通过细调之后的图像, 只是

改变了图像的清晰度, 与预训练结果并没有太大出入.

预训练对深度神经网络训练的成功至关重要, 没有预训练的网络随着层次增加, 容易陷入不好的局部极值点无法出来, 越是低层的网络越需要预训练. 大量研究表明预训练具有非常多的好处, 它具有更好的泛化能力, 它可以帮助网络找到较好的局部极值点等等 [53,54]. 随着训练数据量的增大, 人们发现预训练的作用不再像从前那么重要, 只要调小学习率依然可以学习出很好的深度神经网络. 然而, 没有经过预训练的网络依然不如经过预训练的网络那么容易学习.

3.3 深度神经网络的改进

深度神经网络的预训练和微调都是十分耗时的过程. 为了提升模型的计算速度, 人们从计算机视觉领域借鉴了 ReLU, 用它代替神经网络中的 S 型单元 [35,55], 从而提升了训练速度, 也提升了识别性能. 当然, 这种单元也导致了更加容易快速过拟合, 因此必须配合 Dropout 进行训练.

除了修改网络神经元结构, 采用二阶训练算法代替一阶算法也是一种加速策略. 免赫森 (Hessian Free) 优化算法近年来开始运用于深度神经网络的训练中 [56]. 这种算法既保留了牛顿法的二阶收敛特性, 又免去了大矩阵求逆的麻烦, 更是能够从容应付非满秩条件下的参数求解. 与一阶的梯度下降法不同, 它不需要分小批逐步修改梯度, 因此模型性能不像梯度下降法那样受到批大小的影响, 参数的更新次数也远少于梯度下降法.

不论是一阶训练算法还是二阶训练算法, 均可修改成适合分布式训练的算法. 同步的梯度下降法需要等待所有计算节点返回的结果, 更新完新模型后才能进行下一批计算. 因此导致了许多时间花费在互相等待上. 异步的梯度下降法在每个计算节点提交参数更新时马上修改模型参数, 更好地达到并行的目的 [57]. 从实验效果看, 这种异步的更新方法似乎产生了更好的随机性, 使得模型性能优于正常的梯度下降算法. 但是, 这种异步更新的方式也导致了并行的量无法太大, 通常只适合 4–5 台机器进行并行计算, 太多的机器将导致模型无法收敛. 类似的, 二阶训练算法也可以修改成分布式训练模式. 免赫森算法本身就很适合进行分布式计算, 只需要把梯度计算和曲率计算放到各个计算节点, 再统一汇总到主控机器统一更新模型即可 [35]. 受限内存的 BFGS 算法也可以类似地进行分布式实现 [57].

除了速度的改进, 人们还研究了许多改进识别性能的算法, 主要包括改进模型结构和改进训练算法. 采用部分连接的 CNN 网络代替全连接的 DBN 网络, 性能上提升明显. 在小规模数据库 (TIMIT 数据库) 和大规模数据库 (switchboard 数据库) 上的测试结果均表明 CNN 的有效性 [58]. 与 GMM-HMM 的情况类似, 时序化的区分度训练算法 (MMI, bMMI, MPE, sMBR 等) 应用于深度神经网络后可以获得性能的大幅度提升 [59,60]. 大体上, 基于帧的 sMBR 性能最佳, 而基于网格 (lattice) 的 MMI, bMMI 和 MPE 则性能略差.

3.4 深度神经网络的优秀性能表现

3.4.1 深度神经网络的声学建模能力

深度神经网络一问世便以超强的声学精准度震惊了整个语音识别领域. 2009 年, Mohamed[61] 首次将 DBN 应用到 TIMIT 语音识别, 刷新了当时国际最好的识别记录. 2010 年, Mohamed[62] 通过优化网络权重、优化状态转移参数和区分度训练技术, 将 TIMIT 最佳记录进一步刷新. 随后, Hamid 等人对神网络结构进行了改造, 使得同一特征的隐藏单元权重可以共享, 从而追踪时域上的不变性, 获得了识别性能的进一步提升 [63]. 最近, 邓力等人又发现, 针对不同的频率段采用不同的共享池大小,

表 1 TIMIT 核心测试集性能对比

Table 1 TIMIT core test result

AM	PER (%)
TRIPHONE GMM-HMM DT with BMMI [66]	21.7
MONOPHONE DBN-DNN ON FBANK(8 Layers) [67]	20.7
MONOPHONE MCRBM-DNN ON FBANK(5 Layers) [68]	20.5
MONOPHONE CONVOLUTIONAL DNN ON FBANK(3 Layers) [63]	20.0
MONOPHONE HP-CNN-DNN ON FBANK WITH DROPOUT [64]	18.7
MONOPHONE Deep LSTM RNN ON FBANK [69]	17.7

可以获得更好的效果 [64]. 表 1 显示了最近几年来在 TIMIT 核心测试集上的性能测试结果. 从公开的对比结果可以看到仅用单音子的 DNN-HMM 模型便已经超越了区分度训练的三音子 GMM-HMM 模型, 而模拟人类听觉系统的卷积神经网络又使性能获得大幅度提升.

深度神经网络的性能在大词汇量连续语音识别中也得到了证实. 2011 年, Seide 等人将上下文引入深度神经网络模型, 在 RT03S FSH 任务上, 深度神经网络模型比 BMMI 训练的 GMM-HMM 模型有 33%的性能改善 [65]. 后续的各种实验结果不断证明, 深度神经网络给语音识别带来突飞猛进的发展 [49]. 深度神经网络已经成为最重要的声学建模技术之一.

3.4.2 深度神经网络的语言建模能力

多元文法模型训练算法简单. 它不仅可以预测训练集中未出现过的句子概率, 还可以方便地转换成有限状态机网络, 从而简化识别器的编码设计并提高识别器的解码速度, 因此被广泛应用于语音识别系统之中. 然而, 多元文法模型只考虑指定词之前的 N-1 个词, 因此难以获得长距离约束. 如果增大 N 以引入长距离约束, 又面临着训练数据稀疏和语言模型体积过大的问题. 长短时记忆 (LSTM) 回归神经网络 (RNN) 通过引入遗忘门控制长时间记忆的长度解决了长距离约束问题 [70]. 格雷夫斯 (Graves) 等人利用双向回归神经网络 (BRNN) 与长短时记忆 (LSTM) 回归神经网络 (RNN) 组合拼接成深度神经网络, 取得了 TIMIT 核心测试集性能的进一步突破, 获得了有史以来最佳的识别水平 [69]. 与声学模型构建方法相似, 也有人试图采用前馈型的神经网络构建语言模型 [71], 从结果上看, 深度的前馈网络性能优于单层的前馈网络, 却不如回归神经网络.

多元文法模型的另一个缺点是过分依赖近距离约束导致的泛化能力下降. 采用词语聚类的方法可以增强语言模型的泛化能力. 深度神经网络构造的语言模型 [72] 也可以采用此方法.

采用深度神经网络代替多元文法模型还有一个好处 —— 它可以在几乎不改变模型大小的条件下构建高元语言模型. 对多元文法模型来说, 增加一个历史词将导致模型急剧膨胀. 对深度神经网络语言模型来说, 增加一个历史词只需要在输入端增加一个输入, 模型的规模可以几乎不变. 从实际效果上看, 这种超高元的神经网络语言模型确实超过了 4 元文法模型的性能 [73]. 如何用超大规模语料构建深度神经网络语言模型, 依然有待突破训练速度的瓶颈.

3.4.3 深度神经网络的多语言共享能力

深度信任网络是一种特征优化再提取和模式分类一体的深度神经网络, 除了最顶层具有模式分类的功能, 其他的层次均用于特征优化提取. 底层网络对发音特征提取的过程是以对比散度最小化为目

标的非监督的训练过程. 因此, 利用深度神经网络实现多语言语音识别只需要修改最顶层的结构, 在最顶层安放不同语种的识别器, 所有语种共享底层的特征提取网络. 谷歌的最新研究成果证实了这种识别器的有效性. 其 DNN 模型比 GMM 模型错误率下降了 15%, 而利用 DNN 进行跨语言识别和多语言识别, 性能比单语条件下还略有改善 [74].

3.4.4 深度神经网络的语义识别能力

语义识别和意图理解是对话理解技术中的关键技术, 它有助于消除语音中的歧义, 获得更加准确的答案. 在深度神经网络的启发下, 人们探索了采用深度神经网络进行句子语义分类的方法. 一种通过层层叠加凸网络 (convex network) 获得的深度凸网络被应用于句子语义分类 [75]. 这种网络可以在极端稀疏的输入特征空间中训练出令人满意的语义分类结果.

3.5 深度神经网络引发的新思考

深度神经网络 (DNN) 以其超高的精准度、超强的普适性, 成功地应用于声学建模、语言建模、多语言处理和语义识别等语音识别相关领域中, 不断给我们带来一个又一个惊喜. 从目前的情况看, 随着模型深度的进一步加深, DNN-HMM 的性能还能进一步提升; 随着训练数据的增加, 深度神经网络并没有表现出像 GMM-HMM 那样的快速饱和. 提高识别精度的脚步似乎还没有终止, 但受限于神经网络训练速度较慢的瓶颈, 还需要研究更加快速的适合并行计算的训练算法. 一阶的 SGD 算法, 二阶的 Hessian Free 算法和 L-BFGS 算法的并行训练算法已获得巨大成功, 但推广到更大规模的并行计算依然有难度, 进一步推广更大的并行规模还需要更加复杂的算法设计. 但基本上可以预知, 未来几年内, 并行训练规模从几十台扩充到几百台机器并行训练应该不成问题, 训练百万至千万级别的神经元, 以及百亿至千亿级别的神经网络联接, 已经不太遥远; 通过数据以及学习算法的不断完善, DNN 无疑将进一步改进语音识别的性能.

传统的特征提取算法中, 有不少由于运用了人类听觉感知机理而获得成功的例子. 深度神经网络的研究热潮刚刚开始, 这类听觉感知机理便已经取得可喜的效果. 如果进一步加以应用, 必可获得更优的性能. 通过修改神经网络结构和训练算法, 深度神经网络应该可以更加逼近人类的听觉系统. 这种系统与人的听觉系统相比, 结构上不一定近似, 但功能特点具有一定的相似性. 它可以在一定程度上模拟人类听觉感知的通道无关性, 模拟人类的听觉分离机制, 模拟人类的抗噪机制等, 进而实现类人听觉感知.

4 类人听觉信息处理

4.1 人类听觉系统与类人听觉

人类的听觉系统是一个复杂的系统. 它具有以下特点: (1) 特定的听觉感知范围; (2) 特定的听觉掩蔽现象; (3) 特定的听觉修复机制; (4) 特定的听觉分离机制; (5) 特定的听觉空间映射机制.

人类的听觉系统具有特定的听觉感知范围. 它只能感知特定频率范围和特定响度范围的声音. 人类可以感知的频率范围大约从 20 Hz~20 KHz 区间. 每个人的感知频率范围不尽相同, 同一个人对不同频率的感知灵敏度也不尽相同. 通过对多个人进行多组听觉测试, 人们获得了平均意义上的绝对听闻曲线和痛阈曲线. 在绝对听闻曲线以下的声音, 人们无法感知; 在痛阈曲线以上的声音, 将会导致听觉器官受损失聪. 频率与响度双重指标, 共同圈定了人类听觉感知的范围.

人类的听觉系统具有特定的听觉掩蔽现象. 这种听觉掩蔽主要表现为频率掩蔽和时间掩蔽. 当相邻频率内的两个音高同时响起时, 这两个频率声音将产生两条不同的掩蔽曲线, 如果有一个声音处于另一个声音的掩蔽曲线之下, 则人耳无法感知这个声音. 类似的, 人耳感知的声音在时间轴上也会产生掩蔽曲线, 它可以与后续的声音混响, 修补后续频谱空缺, 从而影响感知内容. 这种影响更多地表现为前面信号对后面信号的影响. 当一帧频谱信号的后一帧信号频谱有残缺时, 前一帧信号可以修复该帧丢失的信息, 使得听觉感知内容依然不变. 反过来, 当 “书” 字的声母 (sh) 的前半段丢失, 则整个语音完全变化成 “猪” 字.

人类的听觉系统具有特定的听觉修复机制. 这种机制一部分表现在听觉掩蔽上, 另一部分表现为频谱的远距离的修复作用. 当声音通过各种窄带信道时, 声音中频率成分被过滤, 其高频成分丢失, 但人耳依然可以正常地对此声音进行识别. 当声音在传播过程中融入周围的噪声时, 频谱的峰值和谷值都受到污染, 甚至在频谱上出现虚假的小峰值, 但人耳依旧可以对说话内容进行正确识别.

人类的听觉系统具有特定的听觉分离机制. 它表现在面对混合声源的声音时, 人类可以选择听取其中特定声源的声音, 把它从复杂的混合声源中分离出来, 典型的如鸡尾酒会问题. 通过双耳接收到的混合声源的信号差异, 人耳可以定位声源和分离声源. 当声源数目小于等于录音通道的数目时, 通过采用独立成分分析法可以分离出多个时间同步的差异信号中的不同声源 [76]. 但是人们已经证实, 人耳拥有更高级的分离机制, 它可以对单通道录音中不同声源进行分离, 实现对特定声源的关注 [77], 或许这种机制来源于对信源本身内部规律的发现, 比如基频的连续性等高级因素 [78].

人类的听觉系统具有特定的听觉空间映射机制. 人耳对音高感知距离是非线性的, 对音位的感知也是非线性的. 无论是对音高感知还是对音位感知, 都是在一个相对空间内部完成的. 不同人的音高不尽相同, 同一个人在不同场景说话音高也不尽相同. 但是无论音高如何变化, 人们都可以正确地感知声调. 类似的, 不同人的共振峰频率也有差异, 通常儿童共振峰最高, 成年男性的共振峰最低, 但这并不影响人们对说话内容的感知.

类人听觉就是模拟人类特定的听觉范围, 特定的听觉掩蔽, 特定的听觉修复, 特定的听觉分离, 以及特定的听觉空间映射, 使得计算机具有与人类相近的听觉能力. 模拟人类听觉系统, 在传统的语音识别系统中获得了许多应用, 这些应用有效地提高了系统的识别性能. 同样, 深度神经网络也可以模拟人类听觉, 从而获得更好的性能提升. 而让神经网络模拟人类听觉系统, 极有可能构造出前所未有的类人听觉系统.

4.2 深度神经网络的类人听觉处理

4.2.1 深度神经网络的听觉抗噪修复机制

深度神经网络具有优秀的抗噪能力. 首先, 得益于深度神经网络对特征的层层结构化抽取, 以及深度神经网络强大的存储能力, 深度神经网络在没有任何修改的条件下便具有可观的抗噪性能. 表 2 对比显示了在平稳噪声和非平稳噪声条件下, 深度神经网络的识别错误率远远低于区分度训练的 GMM-HMM 模型. 与传统的 VTS, PNCC 等抗噪算法不同的是, 传统抗噪算法在提高噪声条件下识别性能的同时, 纯净条件下的识别率略有下降; 而 DNN 在纯净测试条件下、平稳噪声条件下、非平稳噪声条件下识别性能均获得显著的提高.

其次, 人们已尝试通过训练数据的改进改善深度神经网络的抗噪性能. 训练数据的改进方法主要包括: (1) 带噪训练; (2) 抗噪训练; (3) 噪声感知训练 (noise-aware training). 带噪训练法通过将含有不同类型不同信噪比噪声的语音特征送入深度神经网络学习, 使得深度神经网络存储各种噪声条件下的

表 2　不同噪声条件下性能对比

Table 2　Performance test under different noise conditions

AM	Setup	Stationary noise (WER)	Non-stationary noise (WER)
GMM-HMM	1650h,BMMI,SinglePass	13.30%	26.14%
DNN-HMM	1650h,2048x7,SinglePass	7.28% (−45%)	17.76% (−32%)

关键模式, 从而实现抗噪. 抗噪训练法通过将抗噪的结果输入给深度神经网络训练, 使得深度神经网络获得各种抗噪结果的关键模式, 进而达到抗噪的目的. 噪声感知训练法将噪声污染后的语音特征和噪声特征同时输入给深度神经网络, 以期深度神经网络自己学习一种有效的抗噪机制. 微软和剑桥的研究人员对这三种方法进行了对比分析, 非常可惜的是从结果上看, 这三种方法并没有表现出太大的性能差异, 它们在抗噪性能表现不突出 [79].

最后, 改善抗噪性能最好事从训练算法本身入手. 通常 Dropout 算法用于防止模型与数据过度拟合, 这种防止协同自适应现象的算法对抗噪也有很好的效果 [79]. SDAE[52] 则是在训练算法中保证了带噪数据经过 AutoEncoder 编码解码后可以恢复出原始的纯净数据. 由于这种训练算法本身带有抗噪功能, 因此在噪声条件下表现优异.

4.2.2　深度神经网络的听觉缺失修复机制

窄带信号可以看成缺失了高频成分的宽带信号. 可以通过修改深度神经网络的结构, 使得丢失的频谱成分对其他成分的相互影响最小化. 例如, 通过子带划分使得频谱信号的缺失位置集中在少数子带中, 并修改深度神经网络的连接关系, 使得缺失的频谱信息被孤立到一个独立的结构体内, 最后再由多个孤立的结构体的输出产生最终判别结果. 这种方法使得多通道数据融合成为可能, 进而增加了训练数据, 因此, 对任何一种通道来说, 其性能都比原来单通道条件下有所提高 [80]. 当然, 不对结构进行调整, 只是进行混合通道数据融合, 依然也可以获得性能提升 [63].

4.2.3　深度神经网络的听觉空间规整机制

人类的听觉空间规整包括音高空间规整和音位空间规整两个重要过程. 人类听觉音高的非线性规整方法已经比较清楚. 采用梅尔滤波器组对频谱信号进行非线性滤波可以完成这种听觉音高规整. 多次实验结果表明, 这种采用听觉音高设计的原始特征, 优于不考虑听觉音高的对数 FFT 谱特征 [63,64]. 因此, 人们没有必要在神经网络上设计听觉音高规整结构.

音位空间规整机制是深度神经网络结构设计的重点. 早期研究成果表明, 不同人的共振峰频率有一定的差异, 成年男性、成年女性、儿童的共振峰频率依次增高. 这种变化有一定的范围, 当共振峰变化超过指定范围时, 语音的感知将变到另一个音位. 在图像识别中, 人们为了追踪运动着的物体, 允许同一个物体在相邻时刻内有一定的位移变化. 其方法是通过局部卷积来保持物体的局部特性, 通过共享池来允许物体发生少量位移时特征的不变性. 这种允许有少量位移的不变性追踪方法被修改成为只能在特定频域范围内追踪共振位移的方法, 并在 TIMIT 核心测试集上取得当时最好的识别效果 [63]. 人类听觉系统所允许的不同频段的共振峰所发生的位移范围不尽相同, 因此采用分段设计共享池大小的方法比统一设定共享池大小具有更优的性能. 这种特性的运用进一步刷新了 TIMIT 核心测试集的最佳性能纪录 [64].

4.2.4 深度神经网络的听觉空间共享机制

在早期的实验语音学研究中, 人们已经认识到, 每个语言拥有独立的音位空间, 不同语言之间的音位空间是成系统的映射关系. 当一个语言映射到另一个语言时, 部分不区分的音位空间可能会被进一步细分成多个音位, 而部分本来区分的音位空间也有可能合并成同一个音位. 一个语言通常还拥有其他语言所没有的音位. 因此从一个语言到另一个语言的映射, 通常还伴随着音位的增减问题. 不可否认, 国际音标是一种非常有效的通用音子集. 它使得不同语言的同一个音位共享同一个符号. 然而发音词典通常不以严式音标记音, 而代之以宽式音标. 这就导致了同一个音标在不同语言里具有不同的表现. 例如: 英语 speak 和 peak 中的 [p] 音, 便是两个截然不同的发音, 它们分别对应汉语中 b 和 p 声母. 因此, 从英语到汉语的音位映射便成了一个棘手的问题. 当语种繁多的时候, 不同语种之间的数据更难以共享. 深度神经网络的特殊结构给这种问题带来新的解决方案. 由于深度神经网络的底层网络无需标签便可进行预训练, 因此, 各种语言可以混在一起预训练一个共享的底层网络, 只在最顶层网络抽取特定语言所需标签即可完成从共享单元到目标音位映射. 这种多个语言共享底层网络的方法称为共享隐层深度神经网络 (share hidden layer based deep neural network) . 为了防止底层共享网络偏向某种特殊语言, 通常将训练数据随机打乱次序再进行 DNN 训练可以获得更好的性能. Google 已经在这方面作出了很好的探索 [74].

4.3 深度神经网络与类人听觉的未来空间

4.3.1 深度神经网络的结构和训练算法是重点

深度神经网络在没有针对性处理时, 已经具有相当优秀的识别性能. 最新的报道相继指出, 引入听觉感知相关知识后, 深度神经网络的性能还可以进一步优化. 对人耳听觉机理的运用使得深度神经网络的性能不断地改进. 特别是对听觉空间规整的研究, 使得深度神经网络的性能大大提升. 如何采用深度神经网络模拟听觉抗噪修复机制、听觉缺失修复机制和听觉分离机制还有很大的改进空间. 设计出可以模拟人耳听觉抗噪修复机制、听觉缺失修复机制和听觉分离机制的网络结构, 设计出这些网络的相应训练算法, 必将成为新的突破点.

4.3.2 模拟人类听觉抗噪修复机理的可能性

在算法上, 模拟人类听觉进行抗噪修复有理论依据可循. 深度信任网络的预训练以对比散度最小化为目标训练, 这个目标函数本身并不带有抗噪功能. 因此通过输入带噪特征或抗噪特征来训练神经网络很难提升网络的性能. 理论上, 只有优化非监督训练的目标函数, 才能使深度神经网络具有抗噪能力. 深度神经网络底层网络的训练算法, 可以看成是对特征编码再解码的过程, 它要求解码之后的特征与原始特征最相似. 然而, 对抗噪任务来说, 我们希望带噪特征经过编码之后再解码, 可以恢复出纯净特征. 前人已经提出多种方案来实现这一目标. 一种是往纯净数据中添加高斯噪声, 并强迫训练算法学习从带噪特征到纯净特征的恢复过程. 一种是抑制 (随机置 0) 或增强 (随机置 1) 纯净数据中的局部频谱成分, 并强制训练算法学习, 使之拥有采用正常频谱成分恢复异常频谱成分的能力. 为了让被污染的频谱成分更好地恢复到原始的纯净状态, 人们提出对被污染的成分进行权重增强的方法. 通过给被污染区域的重构误差或者重构互信息赋予一个较大的权重, 可以使得算法更好地从正常谱成分中恢复出异常成分的原貌. 上述方法虽然尚未在语音识别领域中应用, 但已经在图像识别领域被证实, 该方法可以使得底层的特征提取网络具有更好的特征选择和重构性能 [52]. 这种算法产生的滤波器具有明显的结构信息, 可以看到, 随着层次的增加, 这些结构化信息越来越有具体的细节. 这些结构化信

息使得神经网络不仅具有高强度抗噪性能, 还具有缺失数据补全功能. 识别结果也证实了其优异的抗噪性能. 噪声的污染源还可以复杂化, 采用非平稳的噪声对信号进行污染, 并强迫特征提取层恢复出纯净的目标. 预训练的准则也可以进一步优化, 例如听觉响度误差最小化准则、听觉信噪比最大化准则等.

在结构设计上, 模拟人类听觉抗噪修复值得尝试. 人类对语音的感知是成时间片的感知, 最好的抗噪算法应该是可以随着时间的推移, 对背景噪声获得越来越准确的估计. 长时间特征在送入识别器之前已经经过了 LDA 等预处理, 而 LDA 所学到的是在无噪声条件下最佳判别分析, 并不意味着带噪条件下依然具有最佳判别性能. 构造新型的网络结构, 使之具有非线性判别分析的降维功能; 构造新型的网络结构, 使之具有长时间噪声估计功能等, 均有待进一步深入挖掘.

4.3.3 模拟人类听觉关注机理的可能性

人类的听觉分离能力在 60 多年前便引起了人们的关注 [80]. 在嘈杂的鸡尾酒会上, 人们可以根据需要选择特定声源的声音进行关注识别. 这一现象引起很多人的浓厚兴趣, 产生了听觉场景分析技术. 各种多通道分离技术 [76] 和单通道分离技术 [78] 也应运而生. 其目标都是要把每个声源单独分开, 以便单独处理. 多通道分离技术要求录音通道数大于等于声源数目, 这种苛刻要求在很多情况下并不现实. 而单通道分离产生的语音质量极差, 很难有很足够高的识别率.

人类的听觉关注机制可以看成对多个声源进行追踪和分离的过程. 多声源追踪可以看成多目标追踪的一种特例. 多目标追踪识别技术在图像识别中已经有广泛的应用. 典型的场景是两个球队踢足球的情况, 每位队员的位置都可能随时发生移动, 但却要求随时追踪每个队员的位置. Bazzani 等人 [81] 提出了一种采用粒子滤波器和深度神经网络相结合的多目标追踪识别法. 这种方法可以时刻追踪目标的运动轨迹, 将目标送给深度神经网络识别. 对语音来说, 多声源追踪与在图像上进行多目标追踪有两个主要不同: (1) 对图像而言, 多个目标通常是分开的, 而且每个目标有自己的运动轨迹; 对语音而言, 多个声源是混合在一起的, 通常多个声源并没有发生相对运动. (2) 对图像而言, 多个目标的叠加是一种相互遮挡覆盖; 而对语音而言, 多个声源的叠加是一种线性或非线性加权. 既然图像上有其特定的目标追踪算法, 语音上应该也可以有类似算法.

另一种解决听觉关注的方法是将多个声源的追踪分离延后到后端识别器中进行处理. 这种方案要求深度神经网络产生多个目标输出. 也就是说, 当两个音 a 和 b 同时响起时, 让深度神经网络在 a 和 b 两处同时产生高概率, 将此概率输出给连续语音识别器, 结合历史搜索结果, 便可实现说话内容自动追踪, 并可对多个声源同时产生解码, 实现多个声源同时识别. 由一个输入特征产生多路输出并不稀奇, 在机器翻译领域已有类似的成功案例. Deselaers 等人 [82] 采用深度神经网络实现了一个音译系统. 该系统将网络的输出分解为 N×M 的矩阵, 对矩阵的每一行求解最佳结果, 便形成了一个具有 N 维输出的多目标识别器. 只需要对最顶层网络的训练算法略加修改, 便可使神经网络具有多个目标输出功能.

在网络结构上, 还可以设计更加复杂的连接关系, 使得声学模型具有显式的目标追踪功能. 例如, 可以在神经网络结构上引入回馈信息, 并模拟维纳滤波的方法进行非线性滤波, 从而达到信息追踪的功能. 当然, 也可以模拟人耳的双通道结构, 通过这种结构使得深度神经网络更好地模拟人耳的双通道听觉感知.

5 结语

语音识别技术经过了半个世纪的积累, 于近年来达到大规模商用的水平. 本文概括了统计语音识别的理论和方法的发展状况, 并单独介绍了深度神经网络的在声学建模、语言建模、多语言共享、语义识别等方面的卓越性能. 深度神经网络的性能优势引起我们强烈的兴趣. 通过回顾类人听觉信息处理对深度神经网络的改进作用, 我们意识到, 深度神经网络与类人听觉信息处理相结合, 必将推进语音识别的进一步发展. 反过来, 深度神经网络技术在语音识别中的进步, 也必将推动类人听觉信息处理技术的进步. 语音识别技术后继发展的重点是对深度神经网络的结构和训练算法的改进使之更好地实现类人听觉, 最后, 我们分析了采用深度神经网络模拟人类听觉的抗噪修复机理和听觉关注机理的可能性.

参考文献

1 Fenn J, Clark W, Natis Y V, et al. Hype cycle for emerging technologies, 2009. Stamford: Gartner, 2009

2 Fenn J, LeHong H. Hype cycle for emerging technologies, 2011. Stamford: Gartner, 2011

3 Uebel L F, Woodland P C. An investigation into vocal tract length normalization. In: Proceedings of European Conference on Speech Communication and Technology, Budapest, 1999. 2527–2530

4 Povey D, Kingsbury B, Mangu L, et al. fMPE: discriminatively trained features for speech recognition. In: Proceedings of ICASSP 2005, Philadelphia, 2005. 961–964

5 Povey D, Kanevsky D, Kingsbury B, et al. Boosted MMI for model and feature-space discriminative training. In: Proceedings of ICASSP 2008, Las Vegas, 2008. 4057–4060

6 Mermelstein P. Distance measures for speech recognition, psychological and instrumental. Pattern Recogn Artif Intell, 1976, 116: 374–388

7 Bridle J S, Brown M D. An experimental automatic word recognition system. JSRU Report 1003: 5, 1974

8 ETSI. Speech Processing, Transmission and Quality Aspects. ETSI ES 201 108, Version 1.1.2, 2000

9 Hermansky H. Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc America, 1990, 87: 1738

10 Woodland P C, Gales M J F, Pye D, et al. The development of the 1996 HTK broadcast news transcription system. In: Proceedings of DARPA Speech Recognition Workshop, Chantilly, 1997. 73–78

11 Hermansky H, Morgan N, Bayya A, et al. RASTA-PLP speech analysis technique. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1992. 121–124

12 Martin R. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2002. I-253–I-256

13 Moreno P J, Raj B, Stern R M. A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1996. 733–736

14 Kim C, Stern R M. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2012. 4101–4104

15 Fan L, Ke D, Fu X, et al. Power-normalized PLP (PNPLP) feature for robust speech recognition. In: Proceedings of International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE, 2012. 224–228

16 Zhan P, Waibel A. Vocal tract length normalization for large vocabulary continuous speech recognition. CMU Computer Science Technical Reports, 1997

17 Emori T, Shinoda K. Rapid vocal tract length normalization using maximum likelihood estimation. In: Proceedings of European Conference on Speech Communication and Technology, Scandinavia, 2001. 1649–1652

18 Pitz M, Ney H. Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans Speech Audio Process, 2005, 13: 930–944

19 Pitz M, Ney H. Vocal tract normalization as linear transformation of MFCC. In: Proceedings of European Conference on Speech Communication and Technology, Geneva, 2003. 1445–1448
20 Uebel L F, Woodland P C. An investigation into vocal tract length normalisation. In: Proceedings of European Conference on Speech Communication and Technology, Budapest, 1999. 2527–2530
21 Haeb-Umbach R, Ney H. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1992. 13–16
22 Zhu M, Martinez A M. Subclass discriminant analysis. IEEE Trans Pattern Anal Machine Intell, 2006, 28: 1274–1286
23 Kumar N, Andreou A G. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Commun, 1998, 26: 283–297
24 Zhu Q, Chen B, Morgan N, et al. On using MLP features in LVCSR. In: Proceedings of International Conference on Spoken Language Processing, Jeju Island, 2004
25 Bourlard H, Morgan N. Connectionist speech recognition: a hybrid approach. Norwell: Kluwer Academic Publishers, 1994
26 Hermansky H, Ellis D P W, Sharma S. Tandem connectionist feature extraction for conventional HMM systems. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2000. 1635–1638
27 Hermansky H, Sharma S. Traps-classifiers of temporal patterns. In: Proceedings of International Conference on Spoken Language Processing, Sydney, 1998. 1003–1006
28 Lee C H, Lin C H, Juang B H. A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans Signal Process, 1991, 39: 806–814
29 Lee C H, Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1993. 558–561
30 Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition. Comput Speech Lang, 1998, 12: 75–98
31 Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang, 1995, 9: 171–185
32 Normandin Y, Cardin R, De Mori R. High-performance connected digit recognition using maximum mutual information estimation. IEEE Trans Speech Audio Process, 1994, 2: 299–311
33 Povey D, Woodland P C. Minimum phone error and I-smoothing for improved discriminative training. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2002. I-105–I-108
34 He X, Deng L, Chou W. Discriminative learning in sequential pattern recognition. Signal Process Mag, Piscataway: IEEE, 2008, 25: 14–36
35 Kingsbury B, Sainath T N, Soltau H. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In: Proceedings of Annual Conference of the International Speech Communication Association, Portland, 2012. 10–13
36 Fry D B. Theoretical aspects of mechanical speech recognition. J British Inst Radio Eng, 1959, 19: 211–218
37 Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2011. 5528–5531
38 Bengio Y, Schwenk H, Sen é cal J S, et al. Neural probabilistic language models. In: Holmes D E, Jain L C, eds. Innovations in Machine Learning. Berlin: Springer, 2006. 137–186
39 Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of ACM International Conference on Information and Knowledge Management. New York: ACM, 2013. 2333–2338
40 Kombrink S, Mikolov T, Karafiat M, et al. Recurrent neural network based language modeling in meeting recognition. In: Proceedings of Annual Conference of the International Speech Communication Association, Florence, 2011. 2877–2880
41 Bellegarda J R. Statistical language model adaptation: review and perspectives. Speech Commun, 2004, 42: 93–108
42 Oba T, Hori T, Ito A, et al. Round-robin duel discriminative language models in one-pass decoding with on-the-fly error correction. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing.

Piscataway: IEEE, 2011. 5588–5591
43 Nguyen P, Gao J, Mahajan M. MSRLM: a scalable language modeling toolkit. Microsoft Research MSR-TR-2007-144. 2007
44 Federico M, Bertoldi N, Cettolo M. IRSTLM: an open source toolkit for handling large scale language models. In: Proceedings of Annual Conference of the International Speech Communication Association, Brisbane, 2008. 1618–1621
45 Hori T, Hori C, Minami Y. Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous speech recognition. In: Proceedings of International Conference on Spoken Language Processing, Jeju Island, 2004. 289–292
46 Fiscus J G. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway: IEEE, 1997. 347–354
47 Zweig G, Nguyen P. SCARF: a segmental conditional random field toolkit for speech recognition. In: Proceedings of Annual Conference of the International Speech Communication Association, Makuhari, 2010. 2858–2861
48 Juang B H, Rabiner L R. Hidden Markov models for speech recognition. Technometrics, 1991, 33: 251–272
49 Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the Shared Views of Four Research Groups. Signal Process Mag, Piscataway: IEEE, 2012, 29: 82–97
50 Hinton G E. Training products of experts by minimizing contrastive divergence. Neural Comput, 2002, 14: 1771–1800
51 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Comput, 2006, 18: 1527–1554
52 Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J Machine Learning Res, 2010, 11: 3371–3408
53 Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning? J Machine Learning Res 2010, 11: 625–660
54 Erhan D, Manzagol P A, Bengio Y, et al. The difficulty of training deep architectures and the effect of unsupervised pre-training. In: Proceedings of International Conference on Artificial Intelligence and Statistics, Clearwater, 2009. 153–160
55 Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
56 Martens J. Deep learning via Hessian-free optimization. In: Proceedings of International Conference on Machine Learning, Haifa, 2010. 735–742
57 Dean J, Corrado G S, Monga R, et al. Large scale distributed deep networks. In: Bartlett P, Pereira F C N, Burges C J C, et al, eds. Advances in Neural Information Processing Systems 25, Lake Tahoe, 2012. 1232–1240
58 Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
59 Vesely K, Ghoshal A, Burget L, et al. Sequence-discriminative training of deep neural networks. In: Proceedings of Annual Conference of the International Speech Communication Association, Lyon, 2013
60 Su H, Li G, Yu D, et al. Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
61 Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition. In: Proceedings of NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, Whistler, 2009
62 Mohamed A, Yu D, Deng L. Investigation of full-sequence training of deep belief networks for speech recognition. In: Proceedings of Annual Conference of the International Speech Communication Association, Makuhari, 2010. 2846–2849
63 Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2012. 4277–4280
64 Deng L, Abdel-Hamid O, Yu D. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013

65 Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks. In: Proceedings of Annual Conference of the International Speech Communication Association, Florence, 2011. 437–440
66 Sainath T N, Ramabhadran B, Picheny M, et al. Exemplar-based sparse representation features: from TIMIT to LVCSR. IEEE Trans Audio Speech Lang Process, 2011, 19: 2598–2613
67 Mohamed A, Dahl G E, Hinton G. Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process, 2012, 20: 14–22
68 Dahl G E, Ranzato M, Mohamed A, et al. Phone recognition with the mean-covariance restricted Boltzmann machine. Adv Neural Inf Process Syst, 2010, 23: 469–477
69 Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
70 Wollmer M, Eyben F, Schuller B, et al. Recognition of spontaneous conversational speech using long short-term memory phoneme predictions. In: Proceedings of Annual Conference of the International Speech Communication Association, Makuhari, 2010. 1946–1949
71 Arisoy E, Sainath T N, Kingsbury B, et al. Deep neural network language models. In: Proceedings of NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, Montreal, 2012. 20–28
72 Hai-Son L, Allauzen A, Yvon F. Measuring the influence of long range dependencies with neural network language models. In: Proceedings of NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, Montreal, 2012. 1–10
73 Schwenk H, Rousseau A, Attik M. Large, pruned or continuous space language models on a GPU for statistical machine translation. In: Proceedings of NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, Montreal, 2012. 11–19
74 Heigold G, Vanhoucke V, Senior A, et al. Multilingual acoustic models using distributed deep neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
75 Tur G, Deng L, Hakkani-Tur D, et al. Towards deeper understanding: deep convex networks for semantic utterance classification. In: IEEE International Conference on Proceedings of Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2012: 5045–5048
76 Hyvarinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw, 2000, 13: 411–430
77 Brungart D S, Simpson B D. Within-ear and across-ear interference in a cocktail-party listening task. J Acoust Soc America, 2002, 112: 2985
78 Hu G, Wang D L. An auditory scene analysis approach to monaural speech segregation. In: Hansler E, Schmidt G, eds. Topics in Acoustic Echo and Noise Control. Berlin: Springer, 2006. 485–515
79 Seltzer M L, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 2013
80 Pylkkonen J. Towards efficient and robust automatic speech recognition: decoding techniques and discriminative training. Dissertation for Ph.D. Degree. Espoo: Aalto University, 2013
81 Bazzani L, de Freitas N, Ting J A. Learning attentional mechanisms for simultaneous object tracking and recognition with deep networks. In: Proceedings of NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop, Hilton, 2010
82 Deselaers T, Hasan S, Bender O, et al. A deep learning approach to machine transliteration. In: Proceedings of EACL 4th Workshop on Statistical Machine Translation, Athens, 2009. 233–241

Some basic problems of speech recognition in the internet era

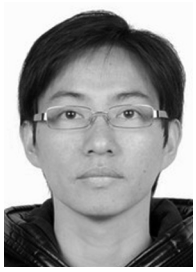
KE DengFeng* & XU Bo

Institute of Automation, Chinese Academy of Sciences, Beijing 100190*, China*

*E-mail: dengfeng.ke@ia.ac.cn

Abstract Speech recognition technology has reached its highly mature period after half a century of development. In this paper, we first summarize the development of statistical theory and methods of speech recognition, and then introduce the excellent performance of deep neural networks in the tasks of acoustic modeling, language modeling, multilingual acoustic sharing and semantic classification. Deep neural networks' performance advances arouse our great interest. By reviewing the pushing effects on deep neural networks of human-like auditory information processing, we realized that the combination of them will promote the further development of speech recognition. In turn, the deep neural network technology for speech recognition will promote the human-like auditory information processing technology as well. Subsequent development of speech recognition technology focuses on the design of the network's structure and the training algorithms of deep neural networks that make it better fit our auditory systems. Finally, we briefly analyze the feasibility of human-like anti-noise processing and human-like auditory attention mechanism implemented by deep neural networks.

Keywords signal processing, speech recognition, neural networks, deep neural network, human-like auditory



KE DengFeng was born in 1980. He received the doctor degree of engineering in pattern recognition and intelligent systems from University of Chinese Academy of Sciences in 2010. He joined Institute of Automation, Chinese Academy of Sciences as a research assistant in 2009, and became associate researcher in 2012. Currently, his research interest includes speech recognition, speech translation, and computer assisted language learning.

XU Bo was born in 1966. He received the doctor degree of engineering in pattern recognition and intelligent systems from University of the Chinese Academy of Sciences in 1997. He has long been engaged in speech and language information processing research. Because of his special contributions to the research field, he has been selected as leading expert of National High-Tech Research & Development Program of China from 2004 to 2010. Currently, he serves as vice director of CASIA (Institute of Automation, the Chinese Academy of Sciences), Vice Chairman of CIPS (Chinese Information Processing Society of China), Director of IDMTech (Interactive Digital Media Technology Research Center) in CASIA. His current research interests include oral speech translation, media and big data intelligent processing, virtual reality technology.