**AI在语音识别行业调查报告**

|  |  |
| --- | --- |
| 学生姓名 | 沈琪瀚 |
| 学号 | 3200607038 |
| 学院 | 计算机学院 |
| 学位类别 | 专业型硕士 |
| 专业领域 | 电子信息 |
| 课程老师 | 马健兵 |

**AI在语音识别行业调查报告**

摘要：AI是当今研究的热点，其中语音识别（ASR)就是AI的感知层面的技术，作为AI的入口，语音识别有着举住轻重的作用。本文通过对语音识别技术的历史、现状、未来发展、相关明星企业、典型的技术应用等进行阐述。在深度学习流行之前，隐马尔可夫模型-高斯混合模(HMM-GMM)是传统语音的主要框架，随着深度学习的快速发展，DNN模块取代GMM,语音识别性能有了一定程度的提升。同时，对于HMM-DNN 模型自身存在很多的不利因素的限制，但是端到端模型(End-To-End ASR)具有简化流程等优点。所以，端到端模型是语音识别未来的重要研究方向。

关键词：语音识别；深度学习；端到端；kaldi；

# 1 引言

随着技术经济的快速发展，互联网和人工智能技术对生活的影响越来越大。人们可以通过键盘、手写输入等与计算机进行通信，但为了更智能地与计算机通信，语音通信显然是一种较好的方法，语音通信的关键技术是语音识别。语音识别起源于1952年，贝尔研究所的实验系统只能识别10个英语数字的发音。在接下来的二十年中，语音识别已经朝着大规模发展的方向发展，但识别的内容仍然十分有限。

在80年代以后，研究人员对语音识别技术的研究方法发生了很大的变化。随着基于统计模型的语音识别理论的成熟，传统的标准模板的匹配技术逐渐被这种新发展出来的方法所替代。同时，为了适应时代的发展，小词汇显然无法满足需求。因此，对大词汇的研究从这一点出发。语音识别的来源不限于特定人的言语。成为研究的焦点。

1956年是国内语音识别技术研究的起点，但当时社会的技术水平和经济条件还不足以支持这一技术的研究。因此，直到国家启动863计划，在充分的经济和硬件支持下，语音识别技术才重新进入科学家的视野，从此开始逐步步入正轨。虽然研究时间不长，但所取得的研究成果对生活的各个方面都有好处，如语音输入方法、智能机器人、智能家居等。

21世纪，深度学习的发展极大地促进了语音识别技术的发展。2017年，微软宣布，其在交换机上的字错误率（WER）已降至5.1%，这意味着在某些条件下，机器已经可以识别出像人类专业速记员这样的单词。2018年阿里巴巴语音实验室，推出的语音识别模型DFSN采用开源框架是基于Kaldi框架构建的，在Fisher（FSH）数据集上的测试字错误率仅为9.4%。百度的模型是在自己建立的中国数据集上进行培训和测试的，而功耗低达7.93%。效果不错。不过，实际的业务场景要复杂的，识别精度会收很多因素的影响，识别率会大大降低。因此，语音识别领域还存在许多问题需要研究和解决。

# 2 Kaldi简介

Kaldi是众多语音识别框架中最常用、社区最活跃的识别工具，目前的维护和支持工作主要由Daniel Povey负责。 Kaldi的代码在github上是开源的，并且在社区上仍然非常活跃。该软件不仅集成了另一种语音识别软件HTK的基本功能，还添加了深度神经网络分类器（DNN）。

对于语音研究人员来说，有许多可以使用的语音识别工具，例如，用C语言编写的HTK，用Java语言编写的sphinx-4等。但是，卡尔迪（Kaldi）具有以下独特优势，使其得以不断开发和应用：

集成了有限状态转换器（FST），以方便开发人员将其用作库。

它支持更广泛的线性代数运算，包括Blas和LAPACK。在其矩阵函数中。

软件和算法的设计采用通用结构，方便开发人员根据自己的需求进行修改和扩展。该软件的代码是开源的，开发人员可以修改代码并重新分发。

Kaldi的核心代码由C ++编写，其上的一些高级工具是使用bash，Python和其他脚本语言编写的。该软件的结构如图1所示。图1中的Kaidi架构主要分为四个部分：第一部分是顶级外部工具，包括上述矩阵操作库；第二部分是第一部分。第二个是Kaldi库，主要包含GMM，HMM和其他训练模型的代码；第二个是编译后的可执行程序；最后一部分是脚本程序，用于控制语音训练的步骤。

目前，作为HTK的正式文件，Kaldi软件的正式文件没有很多有关统计语音识别的基本介绍性文件。因此，对于非专业的语音识别研究人员或语音基础知识较弱的人，他们需要在使用该软件之前充分了解语音信号处理的基础知识。除了上面提到的独特优势之外，Kaldi软件还具有以下对开发人员有利的功能。第一点是对Kaldi的代码进行了全面测试，以确保代码的正确性，不仅可以确保软件的稳定性，还可以方便开发人员快速准确地定位错误。第二点是，Kaldi的代码相对容易理解。了解语音信号处理的基本知识后，我们可以快速了解代码的含义以及每个模块在代码中的作用。最后一点是，Kaldi中的每个模块相对较小，并且每个模块之间的耦合也相对较小，因此修改和重用非常方便。

**3 国内外研究现状**

语音识别技术的研究开始二十世纪50年代。1952年，AT&Tbell实验室的Davis等人成功研制出了世界上第一个能识别十个英文数字发音的实验系统：Audry系统。

60年代计算机的应用推动了语音识别技术的发展，提出两大重要研究成果：动态规划(Dyna[mi](http://www.hqpcb.com/" \t "http://www.elecfans.com/d/_blank)c Planning， DP)和线性预测分析([Linear](http://www.elecfans.com/tags/linear/" \t "http://www.elecfans.com/d/_blank) Predict， LP)，其中后者较好的解决了语音信号产生模型的问题，对语音识别技术的发展产生了深远影响。

70年代，语音识别领域取得突破性进展。线性预测编码技术(Linear Predict Coding， [LPC](http://www.elecfans.com/tags/lpc/" \t "http://www.elecfans.com/d/_blank))被Itakura成功应用于语音识别；Sakoe和Chiba将动态规划的思想应用到语音识别并提出动态时间规整算法，有效的解决了语音信号的特征提取和不等长语音匹配问题；同时提出了矢量量化（VQ）和隐马尔可夫模型（HMM）理论。在同一时期，统计方法开始被用来解决语音识别的关键问题，这为接下来的非特定人大词汇量连续语音识别技术走向成熟奠定了重要的基础。

80年代，连续语音识别成为语音识别的研究重点之一。Meye[rs](http://www.elecfans.com/tags/rs/" \t "http://www.elecfans.com/d/_blank)和Rabiner研究出多级动态规划语音识别算法(Level Building，LB)这一连续语音识别算法。80年代另一个重要的发展是概率统计方法成为语音识别研究方法的主流，其显著特征是HMM模型在语音识别中的成功应用。1988年，美国卡内基－梅隆大学(CMU)用VQ/HMM方法实现了997词的非特定人连续语音识别系统SPHINX。在这一时期，人工神经网络在语音识别中也得到成功应用。

进入90年代后，随着多媒体时代的来临，迫切要求语音识别系统从实验走向实用，许多发达国家如美国、日本、韩国以及IBM、Apple、AT&T、NTT等著名公司都为语音识别系统实用化的开发研究投以巨资。最具代表性的是IBM的[Via](http://www.hqpcb.com/zhuoluye11/?tid=26&plan=fashaoyou" \t "http://www.elecfans.com/d/_blank)Voice和Dragon公司的Dragon Dectate系统。这些系统具有说话人自适应能力，新用户不需要对全部词汇进行训练便可在使用中不断提高识别率。

**4 现状和未来展望**

语音识别主要趋于远场化和融合化的方向发展，但在远场可靠性还有很多难点没有突破，比如多轮交互、多人噪杂等场景还有待突破，还有需求较为迫切的人声分离等技术。新的技术应该彻底解决这些问题，让机器听觉远超人类的感知能力。这不能仅仅只是算法的进步，需要整个产业链的共同技术升级，包括更为先进的传感器和算力更强的芯片。

单从远场语音识别技术来看，仍然存在很多挑战，包括：（1）回声消除技术。由于喇叭非线性失真的存在，单纯依靠信号处理手段很难将回声消除干净，这也阻碍了语音交互系统的推广，现有的基于深度学习的回声消除技术都没有考虑相位信息，直接求取的是各个频带上的增益，能否利用深度学习将非线性失真进行拟合，同时结合信号处理手段可能是一个好的方向。

（2）噪声下的语音识别仍有待突破。信号处理擅长处理线性问题，深度学习擅长处理非线性问题，而实际问题一定是线性和非线性的叠加，因此一定是两者融合才有可能更好地解决噪声下的语音识别问题。（3）上述两个问题的共性是目前的深度学习仅用到了语音信号各个频带的能量信息，而忽略了语音信号的相位信息，尤其是对于多通道而言，如何让深度学习更好的利用相位信息可能是未来的一个方向。

（4）另外，在较少数据量的情况下，如何通过迁移学习得到一个好的声学模型也是研究的热点方向。例如方言识别，若有一个比较好的普通话声学模型，如何利用少量的方言数据得到一个好的方言声学模型，如果做到这点将极大扩展语音识别的应用范畴。这方面已经取得了一些进展，但更多的是一些训练技巧，距离终极目标还有一定差距。

（5）语音识别的目的是让机器可以理解人类，因此转换成文字并不是最终的目的。如何将语音识别和语义理解结合起来可能是未来更为重要的一个方向。语音识别里的 LSTM 已经考虑了语音的历史时刻信息，但语义理解需要更多的历史信息才能有帮助，因此如何将更多上下文会话信息传递给语音识别引擎是一个难题。

（6）让机器听懂人类语言，仅靠声音信息还不够，“声光电热力磁”这些物理传感手段，下一步必然都要融合在一起，只有这样机器才能感知世界的真实信息，这是机器能够学习人类知识的前提条件。而且，机器必然要超越人类的五官，能够看到人类看不到的世界，听到人类听不到的世界。

**5 相关的明星企业**

科大讯飞（002230）：科大讯飞股份有限公司成立于1999年,是一家专业从事智能语音及语言技术、人工智能技术研究,软件及芯片产品开发,语音信息服务及电子政务系统集成的国家级骨干软件企业。

三泰控股（002312）：2019年公司年报披露：子公司维度金融多年来始终将技术创新作为重中之重，积极探索各类前沿新兴技术，推动大数据、语音识别等技术在金融服务外包中的应用，成功降低了服务成本，提升了服务效率，优化了客户体验。

圣邦股份（300661）：2017年9月18日，圣邦股份在互动平台上表示，公司重视人工智能领域模拟芯片产品的开发，积极布局，并已有多款产品应用于人工智能领域，如公司多款产品已应用于智能语音识别、语音回放与交互、传感器测量、超声测距、红外避障等人工智能应用领域。

润欣科技（300493）：2016年公司积极布局射频和功率放大器件、传感器芯片,和欧美、日本的多家射频和传感器芯片厂商合作,设计和销售无线射频和功率放大器件、移动感应、视频识别、指纹和生物识别、语音识别等多种涉及传感、智能计算及应用解决方案,涉及移动支付、智能家居、智能穿戴等多个领域。

四维图新（002405）：公司导航电子地图与动态交通信息业务产品包括导航电子地图产品包括基础地图产品、NDS标准地图格式产品、ADAS高精度地图产品、三维地图数据、语音识别地图产品、行人导航地图产品、室内地图与停车场导航地图产品等。

御银股份（002177）：2017年年报称，公司在生物识别技术方面，开发了多模态生物识别解决方案平台，进一步完善人脸识别、语音识别、虹膜识别、指静脉识别等系统；可以实现自助设备刷脸取款功能，用户可以不带卡，只要刷脸就能取款；语音识别系统，在自助设备上加入语音识别系统，语音识别进行人机交互，可以实现取款、存款、查询等功能。

神思电子（300479）：2018年7月，公司正式成为百度AI生态合作伙伴，依托百度全球领先的人工智能服务平台，在人脸识别、图像处理、语音识别、数据智能、自然语言处理等技术领域，共同打造更具差异化与竞争力的商业人工智能解决方案，加速在安防、金融、通信、医疗、政务等行业商业人工智能场景化落地应用，推进人工智能与实体经济深度融合。

康力电梯（002367）：2014年12月份，公司以现金5330万元增资参股紫光优蓝，其中4000万元用于增资，1330万元用于受让紫光股份所持有紫光优蓝14.285万元出资额，增资转让完成后，公司共持有紫光优蓝40%的股权。

**6 典型的技术应用**

语音识别技术经历了语音识别、语音合成以及自然语音合成3个阶段。从原理上讲,似乎让计算机识别人的语言并不难,其实困难还是不少的。例如,不同的人读同一个词所发出的音在声学特征上却不完全相同;即便是同一个人,右不同情况下对同一个字的发音也不相同。加上人们讲话时常有不合语法规律的情况,有时还夹杂些俗语,或省略一些词语,而且语速变化不定。所有这些,在我们听别人讲话时似乎都不成为问题,但让机器理解则很是困难。近年来,由于计算机功能的日益强大,存储技术、语音算法技术和信号处理技术的长足进步,以及软件编程水平的提高,语音识别技术已经取得突破性的进展,使它的广泛应用成为可能。

语音识别技术的应用主要有以下两个方面。一是用于人机交流。目前这方面应用的呼声很高,因为使用键盘、鼠标与电子计算机进行交流的这种方式,使许多非专业人员,特别是不懂英语或不熟悉汉语拼音的人被拒之于门外,影响到电子计算机的进一步普及。语音识别技术的采用,改变了人与计算机的互动模式,人们只需动动口,就能打开或关闭程序,改变工作界面。这种使电脑人性化的结果是使人的双手得到解放,使每个人都能操作和应用计算机。电话仍是目前使用最为普遍的通信工具,通过电话与语音识别系统的协同工作,可以实现语音拨号、电话购物以及通过电话办理银行业务、炒股、上网检索信息或处理电子件等。不久,能按主人口令接通电话、打开收音机,以及通过声纹识别来者身份的安全系统也将获得应用。

语音识别技术的另一方面应用便是语音输入和合成语音输出。现在,已经出现能将口述的文稿输入计算机并按指定格式编排的语音软件,它比通过键盘输入在速度上要提高2～4倍。装有语音软件的电脑还能通过语音合成把计算机里的文件用各种语言“读”出来,这将大大推进远程通信和网络电话的发展。

**7 调研理由**

语音识别有巨广阔的应用前景，把语音识别合理的利用起来，将会极大的提高工作效率。

**8 技术细节**

语音识别目前在无杂音的条件下，识别已经很高了但是在有噪音的情况下识别率还待提升。并且通过深度学习方式去训练模型的话需要整理不同的数据集，在信号处理上有不同的方式。模型的效果和数据集的大小有很大的关系，在一定范围内，训练数据时长越长训练的效果越好。

**9 结论**

通过以上的分析，我们知道目前语音识别行业对普通话的准确率还是很高的，但是对于特定行业对识别效果还是很差的如：方言识别，噪音较大的识别。所以未来的发展，我们应该注意这些方面。如果想进军语音识别领域我们可以对上述方面引起注意。

**10文献索引**

[1] 朱春山. 基于 Kaldi 的语音识别的研究[D]. 南京:南京邮

电大学,2018.

[2] 何湘智. 语音识别的研究与发展[ J]. 计算机与现代化,

2002(3):3-6.

[3] 胡文君,傅美君,潘文林. 基于 Kaldi 的普米语语音识别

[J]. 计算机工程,2018,44(1):199-205.

[4] 杨胜捷,朱灏耘,冯天祥,等. 基于 Kaldi 的语音识别算法

[J]. 电脑知识与技术,2019,15(2):163-166.

[5] 刘 琼. 几种开源英语识别工具包的对比分析[J]. 计算技

术与自动化,2018,37(4):123-127.

[6]Davis KH, Biddulph R, Balashek S. Automatic recognition

of spoken digits. The Journal of the Acoustical Society of

America, 1952, 24(6): 637-642. [doi: 10.1121/1.1906946]

[6] Davis SB, Mermelstein P. Comparison of parametric

representations for monosyllabic word recognition in

continuously spoken sentences. IEEE Transactions on

Acoustics, Speech, and Signal Processing, 1980, 28(4):

357-366. [doi: 10.1109/TASSP.1980.1163420]

[7]Vintsyuk TK. Speech discrimination by dynamic

programming. Cybernetics, 1968, 4(1): 52-57.

[8]Sakoe H, Chiba S. Dynamic programming algorithm

optimization for spoken word recognition. IEEE Transactions

on Acoustics Speech and Signal Processing, 1978, 26(1):

43-49. [doi: 10.1109/TASSP.1978.1163055]

[9]Buzo A, Gray A, Gray R, et al. Speech coding based upon

vector quantization. IEEE Transactions on Acoustics,

Speech, and Signal Processing, 1980, 28(5): 562 -574. [doi:

10.1109/TASSP.1980.1163445]

[10]Lee KF, Hon HW, Hwang MY, et al. The SPHINX speech

recognition system. International Conference on Acoustics.

IEEE, 1989.