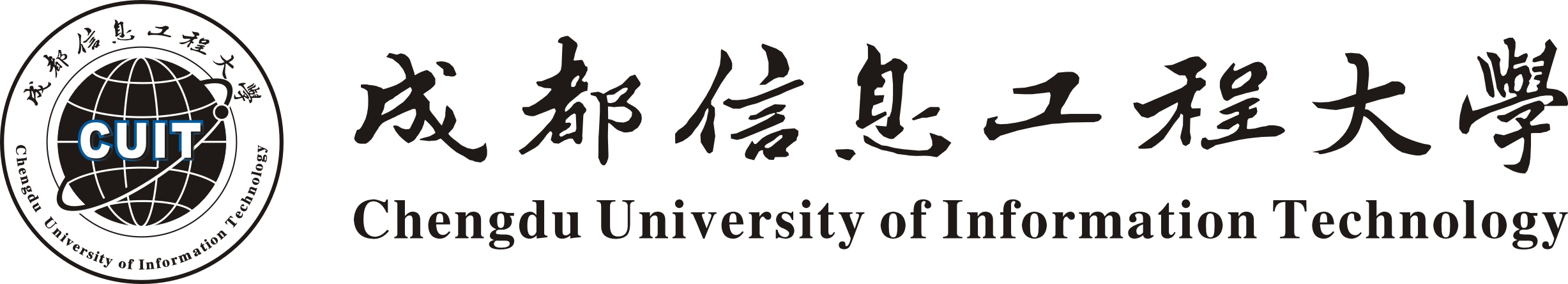
****

**研究生《人工智能》**

**课程论文**

|  |  |
| --- | --- |
| 学生姓名 | 张卓然 |
| 学号 | 3200607039 |
| 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在语音行业调查报告**

摘要： 2016年是人工智能(AI)元年，智能慢慢走近到了人们生活中的方方面面，越来越多的企业也用到了智能应用。智能语音技术就是其中的一部分，智能语音技术作为AI应用最成熟的技术之一，在智能车载、智能家居、智能购物、智能搜索、智能可穿戴、智能教育、智能纠错、智能电话等领域都有了迅猛的发展，因此，现如今越来越多的互联网巨头、科技创新公司、新型创业公司纷纷从不同的维度进入智能产业，从而影响到未来面向物联网智能语音的发展，也会影响到语音商业模式的变化。本文将通过智能语音介绍、国内外现状以及未来发展、典型技术等方面来对语音行业进行一个简单的介绍。

关键词：智能语音、语音识别、语音情感分析、深度学习

# 1 引言

近些年，随着语音技术和人工智能技术的不断发展，智能化与计算机技术不断结合。越来越多的语音技术在深度学习算法的支撑下，获得了更好的效率与精度，机器语音识别准确率第一次达到人类水平，意味值智能语音技术落地期地到来。

语音情感识别是通过分析说话人的语音来识别人的情绪状态。语音情感识别的主要流程包括选择情感特征，情感特征的提取与降维和分类。其中的每一步都具有重要的研究价值。目前公认的与语音情感相关的声学特征有韵律学特征、音质特征、基于谱的相关特征，也有学者提出使用其他特征，如语音的非线性特征。语音情感识别准确率与语音特征的选取和提取有直接关系，对数据降维或者对提取算法做出改进，能有效提高情感识别率。如 Sun 等人提出的一种基于最大 Teager 能量算子的过零特征提取方法可以更好地反映不同情感状态的特征。除了特征外，还需要使用合适的模型对情感进行准确的分类。比较常用的识别模型有隐马尔可夫模型、高斯混合模型、支持向量机和神经网络。

随着计算机技术以及人工智能领域的快速发展，语音情感识别逐渐在各个方面发挥了重要作用，尤其是在人机交互领域。智能机器人可以在与用户进行实时互动时，通过识别用户的情绪采取不同的回应，提升交互间的真实性与趣味性。在呼叫中心，通过嵌入以语音识别为核心的智能质检系统，可以指引运营人员发现服务过程中的短板，提升客户满意度。

智能语音交互是以语音输入为基础的新型交互模式，通过与机器进行语音交流就能够接收到反馈的结果，其系统主要由语音的采集与识别、语义的理解以及语音的合成组成。语音采集主要是对音频进行录入和采样以及对音频进行编码; 语音的识别主要是利用对语音识别的引擎来把采集的语音按照特定的场景与语音识别数据库中的数据进行对比匹配，待匹配成功后，把匹配成功获得的结果转变为文本或者命令进行输出; 语义的理解主要是把语音识别所转出的文本或命令完成对应的操作; 语音合成是将对应的文本或命令转化为声音输

出。语音识别数据库既可以放在终端也可以放在云端上，由于目前所使用的语音识别数据库的算法比较复杂并且数据量巨大，所以大都选择放在云端。语义理解数据库同语音识别数据库一样既能放在终端也能放在云端，由于这个数据库需要与操作系统 或 应 用 进 行 耦合绑定，所以通常会放在终端。

**2 国内外研究现状**

人的声音随着不同的情感表达，有着明显的不同。不同的情绪有着不同的表达，愤怒时声音变大，声调变高，而悲伤时声音总是低沉的。Murray&Arnott ( 1993 ) 指出 高语速与愤怒的情绪有关。

另外，语音情感识别技术在越来越来多的方面展示着自己的异彩。例如：在驾驶员 情感识别方面，Tawari 等人进行了基于语音情感识别的汽车驾驶员辅助系统的研究，该研究的目的在于通过对驾驶员语音进行情绪识别并给出提示信息，使驾驶员在开车过程中能够保持良好的驾驶情绪，进而提高驾驶员的驾车安全性，同时一定程度上也为减少交通事故的发生做出了贡献。

除了在驾驶员情感识别方面的应用外，语音情感识别技术在刑事案件、智能现代化等方面均有着很好的应用。

**3 发展史**

语音识别技术最初的兴起是在 20 世纪 50 年代，此时我国的语音识别技术也开始出现，国外当时只是在贝尔实验室里进行关于识别英文数字的简单发音装置的研究，而后带来了小词汇量和单个词语的识别，在进入20世纪70年代，才有了语音识别实质性的展，出现了逐渐成为主流的隐式马尔科夫模型技术（HMM），并且从传统的目标匹配方式向基于统计的数学化方向有很大的进展。进入 20 世纪 80 年代，语音识别技术继续深入发展，人工神经网络进入模式识别的范畴，出现了利用人工神经网络进行语音识别问题的处理思路，研究不只是在简单词语数字的识别上，有了连续语音的识别和基于大量词语的识别。我国在此时期执行 863 计划，语音识别技术的研究与国外同步发展，这是语音识别技术的一个过渡阶段。20 世纪 90 年代之后，语音识别技术逐渐进入饱和阶段，基于语音识别的产品开始进入人们的生产生活，随着时代的不断迁移，基于 BP 神经网络的优点，大量的语音识别都基于此进行识别系统产品的设计。

最早的真正意义上的语音情感识别相关研究出现在 20 世纪 80 年代中期,它们开创了使用声学统计特征进行情感分类的先河.紧接着,随着 1985 年 Minsky 教授“让计算机具有情感能力”观点的提出,以及人工智能领

域的研究者们对情感智能重要性认识的日益加深,越来越多的科研机构开始了语音情感识别研究的探索.

在 20 世纪 80 年代末至 90 年代初期,麻省理工学院多媒体实验室构造了一个“情感编辑器”对外界各种情感信号进行采集,综合使用人体的生理信号、面部表情信号、语音信号来初步识别各种情感,并让机器对各种

情感做出适当的简单反应;1999 年,Moriyama 提出语音和情感之间的线性关联模型,并据此在电子商务系统中建造出能够识别用户情感的图像采集系统语音界面,实现了语音情感在电子商务中的初步应用[4].整体而言,语音情感识别研究在该时期仍旧处于初级阶段,语音情感识别的研究主要侧重于情感的声学特征分析这一方面,作为研究对象的情感语音样本也多表现为规模小、自然度低、语义简单等特点,虽然有相当数量的有价值的研究成果相继发表,但是并没有形成一套被广泛认可的、系统的理论和研究方法. 进入 21 世纪以来,随着计算机多媒体信息处理技术等研究领域的出现以及人工智能领域的快速发展,语音情感识别研究被赋予了更多的迫切要求,发展步伐逐步加快.2000 年,在爱尔兰召开的 ISCA Workshop on Speech and Emotion 国际会议第 1 次把致力于情感和语音研究的学者聚集在一起.近年来,先后又有若干以包括语音情感计算在内的情感计算为主题的会议和期刊被创立,并得到了世界范围内的注目。

近 10 余年来,语音情感识别研究工作在情感描述模型的引入、情感语音库的构建、情感特征分析等领域的各个方面都得到了发展.Cowie 等人开发的 FEELTRACE 情感标注系统为语音情感数据的标注提供了标准化工具.Grimm等人将三维情感描述模型(activation-evaluation-power space)引入到自发语音情感识别的研究中,并将维度情感识别问题建模为标准的回归预测问题.Grimm 的工作为维度语音情感识别研究的发展争取到更多的关注,激发了维度语音情感识别的热潮.慕尼黑工业大学的 Eyben 等开发了面向语音情感特征提取的开放式工具包 openSMILE,实现了包括能量、基频、时长、Mel 倒谱系数等在内的常用语音情感特征的批量自动提取,并逐渐得到广泛认可.McKeown 等人以科研项目为依托,创建了一个以科学研究为目的的大型多媒体情感数据库 SEMAINE,并提供了情感数据的维度标注结果,为语音情感识别的研究和发展提供了公开的、丰富的、高质量的自然情感语料.正是这些研究成果的不断涌现,为构建语音情感识别标准化平台做出了里程碑式的贡献。

**4未来展望**

智能化作为未来的发展趋势，现如今语音方面主要涉及到语音和语义，而其中包含语音合成、语音识别和自然语言处理（NLP）三项主要技术。语音合成技术发展最早，应用已较为普遍，除了合成音仍偏机械之外，基本不存在太大技术问题；语音识别在2012年卷积神经网络（CNN）应用之后，准确率大幅提升，已经在C端、B端得到了广泛应用，但效果和体验还不够理想；NLP技术虽然在搜索引擎中早有应用，但在人机交互领域仍属于浅层处理。语音识别“鲁棒性“问题显著在生物学中，有个术语叫做“鲁棒性”，是指系统在扰动或不确定的情况下，仍能保持它的特征行为。这一问题在语音识别领域也存在。语音识别整个过程包含语音信号处理、静音切除、声学特征提取、模式匹配等多个环节。由于语音信号的多样性和复杂性，系统只能在一定限制条件下才能获得满意效果。在真实使用场景中，考虑到远场、方言、噪音、断句等问题，准确率会大打折扣。目前业内普遍宣称的97％识别准确率，更多的是人工测评结果，只在安静室内的进场识别中才能实现。要解决语音识别鲁棒性问题，需要在技术和产品两方面进行优化。一方面，在语音增强、麦克风阵列以及说话人分离等多项技术领域持续投入，并结合后端语义，促进对上下文的理解，从而提升识别效果；另一方面，需要从产品设计上进行优化，比如通过进一步交互，使语音识别变得更为准确。语义分析仍是浅层处理NLP技术大致包含三个层面：词法分析、句法分析、语义分析，三者之间既递进又相互包含。

**5相关的明星企业**

云知声：云知声智能科技股份有限公司专注于物联网人工智能服务，是一家拥有完全自主知识产权、世界顶尖智能语音技术的人工智能企业。公司成立于2012年6月，总部位于北京，在上海、深圳、厦门、合肥设有子公司，目前员工接超过500人。

从交互入手，云知声构建了语音感知、认知和表达、超算平台与图像、机器翻译等多模态人工智能硬核技术，并将这些能力封装在自研 AI 芯片之上，通过“云端芯”一体化产品体系面向行业推出全栈式 AI 技术能力，打造从 AI 技术创新到产业应用的生态闭环。

在应用层面，云知声提供跨硬件平台、跨应用场景，端云一体的人工智能整体解决方案，广泛应用于家居、医疗、金融、教育、交通、汽车、地产等领域。迄今为止，云知声的合作伙伴数量超 2 万家，主要客户涵盖平安、世茂、吉利、格力、美的、海尔、华为、京东、360 等头部企业，覆盖用户超 2 亿，云平台日调用量超 5.7 亿次，覆盖设备超 2.5 亿台。

科大讯飞：科大讯飞股份有限公司成立于1999年，是亚太地区知名的智能语音和人工智能上市企业。自成立以来，长期从事语音及语言、自然语言理解、机器学习推理及自主学习等核心技术研究并保持了国际前沿技术水平；积极推动人工智能产品研发和行业应用落地，致力让机器“能听会说，能理解会思考”，用人工智能建设美好世界。

作为技术创新型企业，科大讯飞坚持源头核心技术创新，多次在机器翻译、自然语言理解、图像识别、图像理解、知识图谱、知识发现、机器推理等各项国际评测中取得佳绩。两次荣获“国家科技进步奖”及中国信息产业自主创新荣誉“信息产业重大技术发明奖”，被任命为中文语音交互技术标准工作组组长单位，牵头制定中文语音技术标准。

同时，科大讯飞还获得了以下荣誉：首批国家新一代人工智能开放创新平台、首个认知智能国家重点实验室、首个语音及语言信息处理国家工程实验室、国家863计划成果产业化基地、国家智能语音高新技术产业化基地、国家规划布局内重点软件企业、国家高技术产业化示范工程等。

2016年，科大讯飞发布讯飞翻译机，开创智能消费的新品类，获得消费市场的广泛认可；2017及2019年，科大讯飞连续两届上榜《麻省理工科技评论》全球50大最聪明公司榜单， 2017年首次入榜，名列全球第六、中国第一。2018年，科大讯飞机器翻译系统参加CATTI全国翻译专业资格（水平）科研测试，首次达到专业译员水平。

2019年，科大讯飞新一代语音翻译关键技术及系统获得世界人工智能大会最高荣誉SAIL（Super AI Leader，即“卓越人工智能引领者奖”）应用奖；同年9月，成为北京2022年冬奥会和冬残奥会官方自动语音转换与翻译独家供应商，致力于打造首个信息沟通无障碍的奥运会。10月，在教育部、国家语委的指导下，承建国家语委全球中文学习平台。

科大讯飞坚持“平台+赛道”的发展战略。基于拥有自主知识产权的核心技术，2010年，科大讯飞在业界发布以智能语音和人机交互为核心的人工智能开放平台——讯飞开放平台，为开发者提供一站式人工智能解决方案。截至2020年12月31日，讯飞开放平台已聚集超过175.6万开发者团队，总应用数超过99.6万，累计覆盖终端用户数30.7亿+，A.I.大学学员总量达到44.4万+，以科大讯飞为中心的人工智能产业生态持续构建。

在平台基础上，科大讯飞持续拓展行业赛道，现已推出覆盖多个行业的智能产品及服务，推动在消费者、智慧教育、智慧城市、智慧司法、智能服务、智能汽车、智慧医疗、运营商等领域的深度应用，TO B+TO C双轮驱动成果显现。

科大讯飞在智能语音和人工智能核心研究和产业化方面的突出成绩得到了社会各界和国内外的广泛认可，作为“中国人工智能国家队”一员已形成共识。创新坚守，初心不改，科大讯飞愿携手广大合作伙伴一起：用人工智能建设美好世界！

**6典型的技术应用**

语音识别：从麦克风输入的音频准确、快速、高效的识别出它是哪个文字。

语义理解：进行智能对话，覆盖衣食住行中绝大部分的使用场景。

语音合成：使用语音合成的技术，将一段文本或文字合成语音。

声纹识别：声音就是密码，自动识别说话人的声音进行身份的认证。

离线操控：无需网络环境，命令词特定设置，随时随地，通过语音进行操作。

**7调研原因**

目前虽然说语音行业发展的不错，但是他只是针对部分场景，比如微信里面的语音，但是现实生活中，有很多场景并没有很好的使用语音，比如不同地区是有不同的方言的，所以在这个方面并没有很好的实现方言的语音识别，而且现如今很多岗位都不能进行疲劳工作，所以通过语音进行疲劳检测也是一个很好的例子，但是由于不同行业语料库的不同，所以并不能很好的将现有模型应用到里面，所以音频这个领域大有作为。

**8技术细节**

因为人说话是靠特定场景来辨别说的是什么的，所以人说话的依赖程度取决于，特定人语音识别，还有飞特定的语音识别，然而说话方式的不同也对语音识别方便造成了影响，如孤立词识别。连续语音识别。然而对于不同的场景，语音里面会存在很多杂音，如：航空领域，因为高度的影响，在进行对讲的时候，实际的音频是很杂乱的，所以这方面降噪的处理就必须的引起我们的注意。

**9结论**

通过上面的论述，我对该领域有了一个大概的了解，也对音频这个领域也有了个更深的认识，然后目前我也在进行语音识别和语音疲劳检测的研究，通过这次调查，我发现我能解决当前我的研究难以突破的地方，也发现，在音频这个领域是可以大放光彩，大有作为的。

**11文献**

[1] 潘玉春;徐明星;贾培发;面向情感语音识别的建模方法研究[J];计算机科学;2007年01期

[2] 王敏妲;;语音识别技术的研究与发展[J];微型机与应用;2009年23期

[3]戴礼荣, 张仕良, 黄智颖. 基于深度学习的语音识别技术现状与展望[J]. 数据采集与处理, 2017(2).

[4]张建华. 基于深度学习的语音识别应用研究[D]. 北京邮电大学, 2015.

[5]语音情感识别研究进展分析[J]. 现代计算机, 2020(20):44-47.

[6]叶硕, 褚钰, 王祎,等. 语音识别中声学模型研究综述[J]. 计算机技术与发展, 2020, 030(003):181-186.

[7] 卢阳. 基于CTC方法的端到端语音识别系统研究与实现[D]. 天津工业大学, 2019.

[8]刘剑. 2017年中国人工智能行业分析——智能语音应用篇[J]. 湖南工业职业技术学院学报, 2017(3).