# 引言

### 目的

在当今智能科技快速发展的时代，语音和声纹识别技术的应用已经渗透到生活的方方面面。然而，在嵌入式设备这一领域，由于计算资源的限制，实现高效且准确的语音和声纹识别仍然面临巨大的挑战。本文的研究正是为了应对这一挑战，通过在嵌入式设备上实现综合语音和声纹识别功能，探索在资源有限的环境中提升模型性能的方法。

首先，本研究对于提升嵌入式设备的智能化水平具有重要意义。嵌入式设备由于其便携性和广泛应用前景，已经成为智能设备的重要组成部分。通过在这些设备上集成语音和声纹识别功能，可以极大地增强其智能交互能力，提高用户体验。例如，在智能家居中，通过语音识别控制家电，在金融安全中，通过声纹识别进行身份验证等，都能够带来显著的便利性和安全性提升。

其次，本研究在算法和模型优化方面的探索，对推动语音和声纹识别技术的发展具有重要的学术价值。通过在有限计算资源条件下实现高效识别，本文提出了一系列创新的方法和技术，包括模型压缩、剪枝、量化等。这些技术不仅可以应用于嵌入式设备，还可以推广到其他计算资源受限的应用场景中，进一步推动整个语音和声纹识别领域的技术进步。

此外，本研究的成果还具有重要的商业应用前景。随着智能设备市场的快速增长，具有高效语音和声纹识别功能的嵌入式设备需求也在不断增加。本文通过实际系统的开发和验证，展示了在嵌入式环境中实现高性能识别功能的可行性，为相关企业提供了技术参考和解决方案，有助于推动产品的创新和市场竞争力的提升。

### 语音识别背景

自动语音识别(Automatic Speech Recognition, ASR)，简称语音识别，是人与人、人与机器顺畅交流的关键技术。随着智能通信设备的蓬勃发展，语音识别技术早已转换成产品，并被广泛应用于会议、客服电话、出行驾驶、教育医疗等各种场景。主流的语音识别技术主要包括：基于机器学习的方法 (如GMM-HMM)和基于深度学习的方法(如 DNN-HMM)。但由于GMM-HMM不仅要求帧内元素之间相互独立，而且没有利用帧间上下文信息，致使模型无法充分刻画声学特征的空间状态分布，识别率较低。而DNN-HMM是有监督训练，由于训练数据人工无法标注，所以需要借助GMM-HMM来实现帧与状态的对齐，因此 DNN-HMM模型依然存在一定局限性。在训练架构上，以上两种主流语音识别模型在声学模型、语言模型和发音词典三大组件上都需要单独设计、训练，步骤比较繁琐。而且这种分阶段系统还需 要声学、语言学等专业知识和技术的积累，存在入门门槛高、开发成本高和难维护等问题。近年来，随着计算能力的快速发展，出现了将传统语音识别技术的三大组件融合成一个模型的端到端语音识别技术，实现了语音到文本的直接映射。为解决语音序列和输出序列长度不一致问题，端到端语音识别技术可分为：连接时序分类(Connectionist Temporal Classification，CTC)、循环神经 网络转换器(RNN-Transducer，RNN-T)以及基于注意力机制(Attention)的方法。Wang等人不仅对这三种模型的发展趋势进行了详细总结，而且深入分析了相关技术的优缺点。随着Transformer在机器翻译领域的广泛应用，Dong等人首次将 Transformer 模型架构引入到语音识别领域，进一步提升了语音识别的准确率。谢旭康等人提出了一种 TCN-Transformer-CTC模型 ，通过时序卷积 (TCN)加强 Transformer对位置信息的捕捉能力。尽管 Transformer在捕获长距离上下文信息上具有较大的优势，但提取局部特征的能力较弱。为解决这个问题，Gulati等人提出了 Conformer模型，该模型在Transformer编码器的基础上加入卷积模块，通过卷积捕获局部细粒度特征，同时保留了Transformer的全局表征能力。Burchi等人提出了一种更为高 效的Conformer模型,进一步降低了计算复杂度。Gao等人提出了一种快速并行的 Transformer模型--Paraformer，将模型的解码速度提升了10 倍以上。Peng 等人提出了一种Branchformer模型，进一步研究了局部特征和全局特征的关系以及对语音识别准确率的影响。Radford 等人提出了一种Whisper模型，该模型支持多任务学习，在解码器里通过 引入prefix prompt来支持任务切换，从而实现多种语言到文本的转换。

### 声纹识别背景

声纹识别技术的发展总体来说可以分为四个阶段。其第一阶段可以追溯到上个世纪的三四十年代，1945年，劳伦斯·科斯塔（L.G.Kersta）等人在美国的Bell实验室以语谱图为基础展开实验，并且进行相关匹配研究，从而使“声纹”的理论得以萌芽。随着时间的推移，“声纹”的概念也被广泛熟知。

第二阶段发生在20 世纪40 年代到70 年代，在这个阶段初步建立了声纹识别理论体系，声纹识别技术的研究集中在如何从声音中提取出能够准确反映身份信息的特征参数上。线性预测倒谱系数（LPCC）由BSAtal提出，这种参数的稳定性较高，可以显著提高该技术的精确性。随着数字信号处理技术的发展，实验人员陆续引入了线性预测编码系数（LPC）和LSP 谱系数等间接特征参数。1963 年，贝尔实验室的S.Pruzansky及其团队开发出一种新型的声纹识别技术，该技术采用模板匹配（Template Matching），并且运用了方差分析的方法，这一创新成果迅速引发了行业的广泛关注。1969年，Luck JE 在声纹识别中首次使用倒频谱技术，并将说话人确认问题作为二分类问题进行实验，最后获得了不错的识别效果。

第三阶段发生在20世纪70年代到80年代末之间，在此阶段，实验者们致力于深入挖掘数字信息的内涵，有助于更加有效的建立模型，同时也在探索新的模式匹配方法，从而大大提升了模型建立的准确性与效率。1971年，BS Atal开始将线性预测倒谱系数（Linear Prediction Cep-strum Coefficient，LPCC）应用在声纹识别领域中，从而大大改善了该技术的精度和可靠性。1972 年，Doddington及其合作伙伴首次尝试利用共振峰参数来实现声音辨认，从而开创性地推动了一种全新的声音辨认技术。这一年，BS Atal也进一步发展了基于基音轮廓的声纹识别技术。

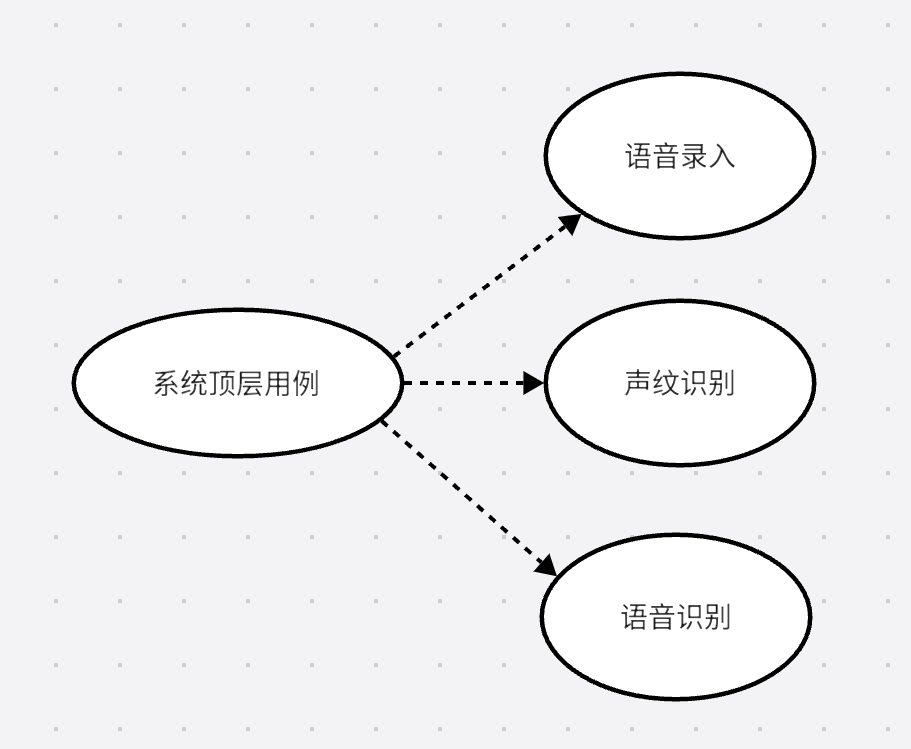
第四阶段发生在20世纪90年代到现在，声纹识别领域的研究人员正在更加关注从声音中提取出关键的特征参数。在1999年，Vergin R和他的团队首次提出了梅尔频率倒谱系数（Mel Frequency Cep-strum Coefficient，MFCC），这一发现为声音识别带来了重大突破。随着互联网技术的不断进步，声纹识别领域引入了许多先进的技术，如动态时间规整（ Dynamic Time Warping ，DTW ）、矢量量化法（Vector Quantization，VQ）、隐马尔可夫（Hidden Markov Model，HMM）、人工神经网络（Artificial Neural Network ， ANN ）等，进一步提升了声纹识别的准确性。D.Reynolds等人提出了高斯混合模型（Gaussian Mixture Model，GMM），在各个领域有着广泛的应用。到了21世纪，P.Kenny和N.Dehak等人研究出了基于联合因子分析（Joint Factor Analysis，JFA）方法，之后又提出了基于i-vector的技术，促进了声纹识别系统性能的提高。在声纹识别领域，深度学习被提出来之前，它们一直是主要方法。

深度学习已成为人工智能领域中的一项重要技术，其对于人工智能的发展和应用有着重要的推动作用。未来，随着人工智能应用的不断拓展和深度学习理论的不断完善，深度学习将在更多的应用场景中发挥其巨大的潜力和价值。2014 年，谷歌公司的VarianiE 及其团队开发出一种基于DNN的声纹识别技术，该技术利用全连接神经网络来提取语音信号的特征，并将每一帧信息输入到深度学习网络中，通过多层处理，可以获得每个说话者的深度说话人嵌入，这些值来自于隐藏层的平均值，这一技术对于基于深度学习的声纹识别具有重要的参考价值。2017年，x-vector系统由D.Snyder及其研究小组提出，该系统利用时延网络技术，将声学特征映射到池化层，声学特征是帧级别的，池化层是语句级别的，从而学习到语句级的深度说话人嵌入，以实现对说话人的识别。NagraniA和牛津大学团队研发了VGGVOX系统，该系统由孪生网络和对比损失来建立，在研究方面获得了不错的成效。百度团队研发了DeepSpeaker系统，该方法利用深度残差网络和循环神经网络提取语音信号的特征，用平均池化获取说话人的嵌入特征，通过交叉熵损失预训练和三元组损失微调进行声纹识别的训练，最后通过打分融合方法提高声纹识别的准确率。Hu及其团队研究出了SE-Net（Squeeze and Excitation Network，SENet），利用SE 块的概念，可以有效地识别和调整不同的通道，从而改变其中的信息传输模式，并且能够有效地控制信息传输。在2018 年，谷歌公司的W.Li 等人引入了与众不同的声纹识别方法，该方法利用GE2E损失函数，将最新值和语音库中所有人一一对比，从而加速模型训练速度，同时也提高了准确率。2021年，肖等人提出在改进ResNet34模型上使用AAM-Softmax 损失函数，针对ResNet34作为背景模型时输入语谱图维度过大以及泛化能力差两大问题做出改进。2022 年，荣等人利用知识蒸馏技术来处理ResNet的数据，并利用深度学习技术来处理这些数据。这种技术通过用蒸馏损失MSE对ResNet声纹特征和I-Vector的差异进行约束，并且能够更精细地分析数据。经过两种独特的数据增强方式的引入，使得该模型能够更好地抵御外界的噪音干扰，从而大大改善了其稳定性和可靠性。此外，该实验也显示出，此技术能够更好地支持声音识别的需求。

# 功能需求分析

### 总体功能需求

#### 系统顶层用例图

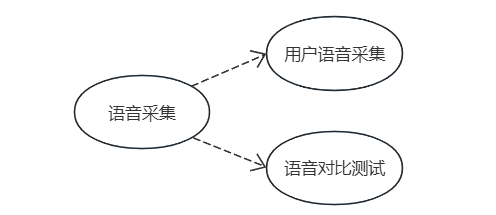


#### 功能需求列表

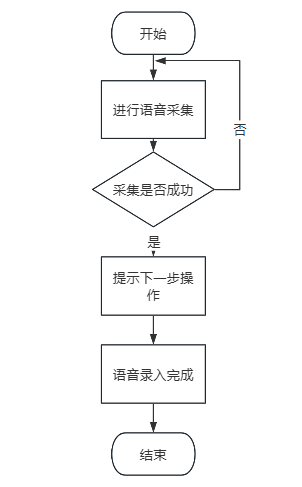
|  |  |  |
| --- | --- | --- |
| 功能类别 | 子模块 | 子功能或需求 |
| 语音采集 | 用户语音采集模块 | 语音录入 |
| 语音对比测试 |
| 声纹识别 | 身份认证模块 | 开机身份认证 |
| 指令识别模块 | 加密身份认证 |
| 语音识别 | 语音内容识别 | 具体指令识别  文字转写 |

### 语音采集

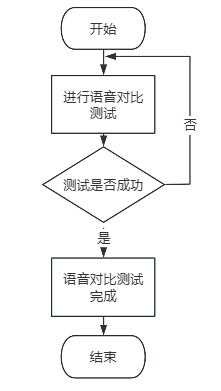
#### 语音采集用例图



#### 语音采集步骤

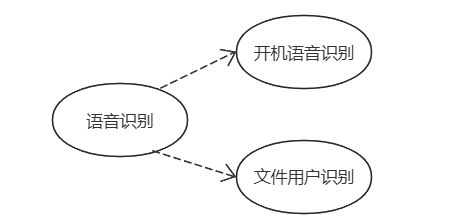


#### 语音对比测试

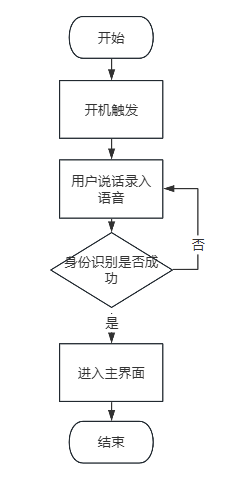


### 语音和声纹识别

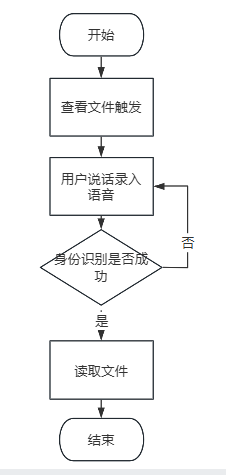
#### 语音和声纹识别用例图



#### 语音识别步骤



#### 文件识别步骤



# 性能需求

### 响应需求

响应时间必须满足如下需求：

安全认证模块响应速度≦6秒（待定）

### 可靠性需求

系统可靠性应满足如下需求：

系统安全可靠性指标达99%，并且具有本分容灾能力

### 可用性需求

系统应满足如下可用性需求：

可对单个用户的多种语言特征进行提取及识别

### 精度需求

语音特征识别率达98%以上，错误拒绝率2%，错误接受率0%

# 非功能需求

### 可用性需求

#### 易用性需求

语音识别系统应具备较好的已用性，各子系统的功能应易于完成，系统界面应比较友好。

### 操作需求

#### 可移植性需求

语音识别系统需要考虑可移植性的需求。具体体现如下：

系统采用Python语言开发

数据库暂时拟用sqlite3数据库

采用嵌入式平台

#### 系统安装需求

系统应支持向导安装，不能有过多的比较专业化的安装配置过程。

#### 可维护性需求

对系统的安装可维护性需要满足如下需求：

若软件因某些原因导致的数据损坏，手机数据库的定时备份能够对损坏数据进行恢复，使得数据能够重新存在新的系统中去。系统的管理操作不能太复杂，应简单易于理解，使维护人员能够快速理解并进行相关的维护操作。

尽量减少系统运行过程中对配置文件进行修改。

系统设计寿命为3-5年。

### 安全需求

#### 系统安全

语音识别系统自身的设计必须满足如下安全性需求：

所选操作系统的安全，可以通过操作系统补丁、防病毒等方式保障系统的安全。

系统必须有较强的身份认证手段，登录系统的人必须经过身份认证才能进行有关的授权业务操作。

#### 文件安全

在进行文件识别时，只有身份认证通过，才可以进行相关的授权操作。