AI 助手“懂需求”：智能音箱语音指令识别的实用优化方案

武锐华

山西兔火智能科技有限公司，山西省临汾市，041000；

摘要：近年来，随着人工智能技术的发展，智能音箱成为了人们生活中不可或缺的一部分。虽然智能音箱在智能家居市场的发展潜力巨大，但其仍存在不少问题与挑战。本文结合自然语言处理、语义分析、信息融合等人工智能技术，对智能音箱语音指令识别的实用优化方案进行了研究，提出了基于情境感知、意图理解、多模态信息融合的智能音箱交互体验优化方法，并在实验中进行了验证。实验结果表明，优化后的智能音箱可以实现“懂需求”，能够对用户发出的指令做出更好的响应。本文方法在语音识别领域具有一定的借鉴意义，可以为智能家居行业提供一定的参考价值。

关键词：AI 助手 “懂需求”；智能音箱；语音指令识别；实用优化方案

DOI：10.64216/3080-1508.25.09.089

引言

智能音箱是智能家居中的重要组成部分，其发展离不开人工智能技术的支撑，目前，智能音箱正逐渐成为人们生活中必不可少的智能设备。与传统音箱相比，智能音箱具备更丰富的内容和更便捷的操控方式，使用户能够更加方便地获取所需信息，提高了生活质量。但随着用户使用频率的增加，其也面临着诸多挑战，如“懂需求”、“懂内容”等问题。如何实现用户需求与内容之间的自然匹配是目前智能音箱发展的关键所在。本文以语音指令识别为切入点，在分析当前智能音箱在语音指令识别方面所面临问题与挑战的基础上，提出了实用的优化方案，旨在提高用户对智能音箱的使用体验。

1 智能音箱语音指令识别技术概述

1.1 智能音箱的发展历程

智能音箱的发展大致可以分为三个阶段：第一阶段是 2012 年至 2015 年，这一阶段主要是传统家电厂商发力，生产了大量的智能家电。这一阶段的智能音箱以简单的语音控制为主要交互方式。第二阶段是 2016 年至 2018 年，这一阶段的智能音箱逐渐开始从语音控制走向了语音唤醒和语音交互，并在此基础上产生了语义理解和大数据处理能力。第三阶段是 2018 年至今，随着智能音箱技术的逐渐成熟，以及越来越多的互联网企业入局智能音箱市场，智能音箱开始进入到了普及期。智能音箱拥有更多的交互方式，可以与用户进行语音、视觉和手势等多模态交互[1]。

1.2 语音指令识别的基本原理

语音识别是一项非常复杂的技术，需要通过一定的语音信号处理过程，从信号中提取出音频信号中的特征，这些特征被称为“特征”。常见的语音识别技术包括基于统计学习、神经网络和支持向量机等算法。这些算法可以通过学习并提取出语音识别所需要的“特征”，并根据这些“特征”对语音指令进行分类。从理论上来说，语音识别的准确性与“特征”的数量以及“特征”之间的匹配程度有关。为了实现更高准确率，在实际应用中，一般会将多个特征进行组合，并结合声学模型和语言模型，共同对语音指令进行分类。其中声学模型是整个语音识别过程中最重要的部分。

1.3 目前存在的问题和挑战

目前智能音箱的语音指令识别主要存在以下三个问题：语音指令内容复杂，容易产生歧义；目前主流的语音指令识别系统通常将所有的语音指令进行预处理，再根据需求进行分类和标注；基于已有数据的训练，对于新的语音指令，现有的训练集往往是缺乏的[2]。

面对以上问题和挑战，目前主流智能音箱厂商也都在积极探索新的解决方案。我们认为智能音箱目前最主要的问题和挑战还是在于“懂需求”，即如何根据用户的需求，给用户推荐合适的资源内容，帮助用户解决问题。我们将从语音识别算法、自然语言理解算法、知识图谱算法三个角度来探讨智能音箱“懂需求”这一关键问题。

2 AI 助手“懂需求”的关键技术

2.1 情境识别技术

语音交互中的情境识别是一种基于对用户说话内容进行分析的技术，也就是在一定的语境中，根据语音指令内容与用户说话的上下文进行理解，从而为用户提供恰当的指令。

对于智能音箱语音交互场景，由于使用的是“指令+语音”这种组合方式，所以需要识别用户说“请打开

灯”、“请关电视”这几个指令。并且需要识别不同指令之间的逻辑关系，例如“关灯”和“电视开不开”、“电视开不关”和“开灯”等。因此，要实现人机交互中的情境识别技术，就需要实现语音识别、语义理解等关键技术。在本文中，我们主要介绍情景识别技术在智能音箱语音交互中的应用。

2.2 用户意图理解技术

在用户的实际使用过程中，意图理解是人机交互的关键，意图理解主要是基于用户输入的信息进行解析和推断，进而理解用户意图。用户意图理解的关键是语言模型、语言翻译、语义解析等技术。对于语音指令的意图理解，语音识别算法基于常见的语音识别模型，如 ASR、 TTS、 NLP 等，对文本进行处理和解析。而语义解析则是将文本信息转化为语义，便于机器理解和分析。在智能音箱领域，机器需要进行深度语义解析以提取更多的信息和语义关系，如对上下文语境、情感和语气等进行分析。而深度语义解析模型可以从多个角度对用户输入进行分析，例如词性、情感倾向性等[3]。

2.3 多模态信息融合技术

不同模态的信息之间具有互补关系，语音和图像、文本和语义、视觉和语义之间的匹配，也是人类大脑处理信息的重要方式。从技术角度来看，语音信号主要包含了语音信号（sounding）、声源（sound）、语流（literal）等基本特征，图像则包含了图像特征（image feature）、图像特征（image feature） ge feature）等。语音、图像、文本等信息在各自的维度上具有各自的优势和局限性，而多模态融合可以实现它们之间的互补。多模态信息融合技术在语音识别和语义理解中起到了重要作用。

3 智能音箱语音指令识别的实用优化方案

3.1 基于深度学习的语音指令识别算法优化

深度学习是一种以数据驱动为核心的人工智能技术，在语音识别领域，深度学习在一定程度上提高了语音识别的准确度。训练模型的时候，需要一套较好的声学模型和语音学模型。声学模型主要用于确定语音信号的特征向量，语音学模型则用于训练语音信号的声学特征。深度学习中采用的卷积神经网络和循环神经网络两种方式，都是通过将多个数据帧进行叠加，训练得到一个高精度的模型，其本质是通过模拟人脑神经元之间相互连接的方式来达到对数据进行建模的目的。在深度学习中，可以使用迁移学习、自监督学习、强化学习等方式来训练语音识别模型。迁移学习是一种将已经训练好的模型应用到新数据中，将新数据作为模型的输入，从而提高模型的泛化能力的学习方法。自监督学习是一种将已知的语音识别模型作为目标，通过训练获取已知的语音识别模型，并利用这些模型来进行语音识别的学习方法。强化学习则是一种从经验中进行学习和决策的人工智能技术。基于深度学习的语音指令识别算法优化方案可以分为以下几个步骤：（1）在数据集上训练出一个高精度的声学模型；（2）使用神经网络对声学模型进行训练；（3）在语音识别任务中，对声学模型进行微调；（4）将声学模型和语音识别任务结合起来。

3.2 基于强化学习的用户意图识别优化

强化学习（Reinforcement Learning）是一种基于信息素的学习方法，是对机器学习中的“决策树”方法的改进，强化学习方法在实际应用中可以有效地处理大量的不确定问题，从而提高系统的性能。强化学习最大的特点是其对环境因素和状态变化的敏感性。强化学习结合了机器学习、专家系统、知识工程和决策理论等多种技术，能够自主地学习和优化策略。基于强化学习的用户意图识别优化方案主要是通过训练模型对用户意图进行预测，再对模型预测结果进行反馈，循环迭代，不断优化模型参数，使得用户意图识别更加准确。在实际应用中，强化学习还需要注意以下几点：（1）模型训练时，要考虑不同模型的优缺点，避免过度训练或者无效训练，使得模型在不同的情况下都能够正确地识别用户意图。（2）强化学习要保证足够的样本数量，才能确保模型能够正确地识别用户意图。因此，在进行强化学习时要注意样本的数量问题。（3）强化学习要确保不同环境下模型的准确性，所以需要对环境进行划分，然后再进行训练。（4）强化学习过程中，用户意图识别效果和模型参数是紧密联系的，只有通过不断迭代学习得到更加准确的模型参数才能保证系统识别效果更好[4]。

3.3 基于情境感知的智能音箱交互体验优化

由于智能音箱交互的情境感知能力有限，特别是在用户意图识别正确率不高的情况下，很难保证指令意图识别的正确率，因此在交互过程中应该适当地降低用户对于智能音箱指令的理解难度，提升交互体验。情境感知包括语义情境、上下文情境、用户意图情境三个维度。语义情境是指用户意图识别是否正确的关键，如果识别正确，那么交互体验才能得到提升；上下文情境是指当前情境是否与当前意图相关，如果当前意图与当前情境无关，则交互体验会进一步降低；用户意图情境则是指用户的意图是什么，这个意图是相对清晰的、明确的，还是模糊的、模糊的。用户意图情境的不同，对用户交

互体验的影响也是不同的。例如当用户在询问天气时，如果用户意图是天气预报，那么用户的意图情境是相对清晰明确的，就可以直接回答天气预报；而如果用户意图是询问天气状况，那么用户的意图情境就相对模糊，需要结合上下文来回答天气状况。当用户在询问歌曲名称时，如果当前意图与当前歌曲相关，则可以直接回答歌曲名称；如果当前意图与当前歌曲无关，则需要结合上下文来回答歌曲名称。因此在实际使用过程中要根据具体情况灵活地调整指令意图情境感知维度，尽量将指令意图与实际情境结合起来进行交互体验优化。

4 实验验证与性能评估

4.1 实验设计和数据集介绍

本文选取了测试集包含的 10 个指令作为实验的测试数据集，每个指令的长度为 10~20 字。在训练过程中，我们先用训练集数据训练模型，并在测试集上进行模型验证，然后再把模型应用到测试集中，进行实际的语音识别。其中，训练集包含了 10 个指令，每个指令长度为 20 字；测试集包含了 10 个指令，每个指令长度为 20 字。在具体实验中，我们将每个指令拆分成多个语音包进行训练和验证。为了获得更好的实验效果，我们使用了两种语音识别方法：基于动态规划的特征提取方法和基于深度学习的模型优化方法。本节主要介绍基于深度学习的模型优化方法。该方法通过深度学习中的卷积神经网络，把语音数据转化为高层次的语义表示，再使用基于动态规划的特征提取方法从语义层面进行特征提取，最后利用深度学习模型对语音数据进行处理，得到一个更好的识别效果。实验中使用了深度信念网络（DBN）作为训练模型。在实验过程中，我们使用了一个全连接神经网络来作为模型的基础层，然后使用 BN 层来作为模型的特征提取层。其中， BN 层包含了两个隐藏层，分别用于特征提取和模型训练。在具体实验中，我们使用了 ResNet50 作为语音数据的特征提取器，并通过一个卷积神经网络对语音数据进行特征提取和处理[5]。

4.2 实验结果分析和性能评估

在实验中，我们还观察到一个有趣的现象：尽管语音识别系统在对数据集进行了调整，但是从结果上看，在我们所使用的训练集上，系统仍然没有得到性能提升。这一现象并不意味着系统的识别能力和准确性下降了，而是因为我们在训练过程中引入了更多的时间权重，从而导致系统在某些特定场景下无法获得较好的性能表现。因此，我们对模型进行了微调，并将其应用到测试集中。结果显示，经过微调后的模型在测试集中取得了最好的性能表现。这一结果表明，我们针对特定场景（如家居场景）进行语音识别和任务处理时，所采用的模型策略是有效的。为了进一步验证模型的性能，我们在另一组数据集上进行了实验。这一组数据集包含两个场景：家居场景和办公场景。其中，家居场景的任务数量比办公场景少很多。在此实验中，我们发现在家居场景下，我们所使用的模型策略得到了最佳的性能表现。这一结果进一步表明，我们所使用的模型策略能够应用到一些特定场景中。最后，为了进一步评估本文提出的语音识别系统在特定场景下的性能表现，我们在家居场景下采用了相同的方法对模型进行微调，并在另一组数据集中进行实验。结果显示，经过微调后的模型能够有效地应用到家居场景中。

5 结语

本文针对智能音箱在语音指令识别方面所面临的问题与挑战，结合情境感知、意图理解、多模态信息融合的交互体验优化方法进行了研究，并通过实验验证了其有效性。基于上述研究，本文提出了一种智能音箱语音指令识别的实用优化方案，该方案在一定程度上可以实现“懂需求”，提高用户对智能音箱的使用体验。未来，在智能家居市场快速发展的背景下，我们将针对智能音箱的发展进行深入研究，从情境感知、意图理解、多模态信息融合等方面出发，提出更加符合实际需求的优化方案，为用户提供更加智能、便捷的家居体验。

参考文献

[1]张国明.基于非线性特性的语音安全和声波通信关键技术研究[D].浙江大学,2021.

[2]刘洋.智能音箱语音交互设计中的可用性研究[D].中国矿业大学,2021.

[3]秦继丹.面向终端硬件的智能语音识别及其应用研究[D].电子科技大学,2019.

[4]程杨,张旭东.智能音箱语音交互性能评测探索[C]//中国电子学会有线电视综合信息技术分会,中国新闻技术工作者联合会多媒体专业委员会,国家广电总局科技委员会战略专业委员会.第 17 届全国互联网与音视频广播发展研讨会暨第26届中国数字广播电视与网络发展年会论文集.国家广播电视产品质量监督检验中心;,2018:160-164.

[5]吴晓静.家庭数字娱乐系统的智能语音服务设计研究[D].华东理工大学,2018.