第 1 章　绪论

1.1　连续语音识别的研究背景和意义

在现代信息化社会中，人与人之间的交流方式呈现多样化，其中语言交流是一种基本方式。随着科技的不断发展，机器不断实现智能化，让机器理解语义并作出相应的决策，具有广泛的应用前景。语音的本质是一种声学特征，是一种最自然的交流方式。众所周知，语音在日常生活当中应用广泛，不仅传播速度快，而且在多种复杂环境下都可传播。在信息化时代的今天，机器在人类生活当中扮演着重要的角色，将语音技术进行实用化，将语音识别等多种技术应用到日常生活当中，对于提高人们的生活质量有着重要的意义[1]。

语音识别的研究有着无穷的魅力，不仅在于它涉及多个研究领域，而且语音识别有着一定的难度与挑战，在近百年中，众多学者在语音识别领域展开了深入的研究。简单来说，语音识别的目的就是为了实现机器智能化，使机器能够听懂人类的语义，甚至能够完成人们赋予它的任务。对语音识别进行研究，首先就需要建立数学模型，经过诸多学者的研究以及语音识别技术的发展，逐渐衍生出了声学模型为核心，语言模型起约束作用的语音识别系统。在连续语音识别的研究当中，HMM 是一种识别效果不错的主流算法，常常用于声学建模。20 世纪 90 年代，Rose 等提出了高斯混合模型[2]，核心思想是用多个高斯来描述 HMM 状态的输出概率分布。90 年代，掀起了语音识别研究的热潮，主要得益于基于 GMM-HMM 声学模型的区分性训练准则和模型自适应方法的提出[3]。直到 21 世纪初，在深度学习大量用于语音识别之前，GMM-HMM 作为一种经典模型一直被广泛应用于研究和产业应用，有着较好的识别率。

近年来，随着 CPU 的发展和大数据时代的到来，计算机的计算能力逐渐提高，带动了语音识别研究的发展，深度学习逐渐应用到语音识别领域。在多个行业实现了实用化，并且市场上的语音技术产品有着较好的性能。其中，苹果的 Siri、亚马逊的 Alexa、讯飞语音输入法、叮咚智能音箱等都是其中的典型代表[3]。

虽然就目前的研究情况来看，基于深度学习的声学模型逐渐成为主流的声学模型，比如 DNN-HMM 声学模型，但 DNN-HMM 模型是以 GMM-HMM 模型为基础的，它的发展仍旧离不开 GMM-HMM 模型。因此，对 GMM-HMM 进行深入研究仍具有重要意义。

由于语音识别过程中涉及的算法较为复杂，为了让从事语音识别的研究人员能够迅速搭建一套语音识别系统，许多机构着力于开发一款集语音信号预处理、特征参数提取、模型训练以及识别解码于一体的工具。其中最为著名、应用最为广泛的就是 HTK。HTK 是由剑桥大学工程系开发，先后经过剑桥大学、Entropic 公司及 Microsoft 公司的不断完善和改进，可在 UNIX/Linux 和 Windows 操作系统上使用。通过调用一系列命令函数，根据需要进行选择，就可以搭建起语音识别系统。本文的实验就是在 HTK 平台上，搭建一个小词汇量的连续语音识别系统，通过调整参数，增加高斯混合个数来提高识别率。

1.2 国内外研究现状

随着社会的进步和科技的不断发展，机器逐渐实现智能化，机器在人类社会生活中发挥着重要作用。随着计算机性能的不断提升以及人机交互的提出，对于语音识别的研究逐渐成为热门。语言可以说是最自然、最重要的交流方式，语音作为语音的声学表现形式，对于语音识别的研究是实现人机交互的一个重要途径。

语音识别技术（又称自动语音识别，Automatic Speech Recognition，ASR）起源于 20 世纪 50 年代。1952 年贝尔实验室的 Davis 等人研制了特定人孤立英文数字识别系统，该系统能识别 10 个英文数字发音。1959 年 MIT 林肯实验室的 Forgie 等人，研发出可识别 10 个元音的语音识别系统。20 世纪 60 年代到 70 年代，学者们先后提出学者们先后提出信号线性预测编码（Linear Predictive Coding，LPC）技术[4]和动态时间规整（Dynamic Time Warping，DTW）[5]技术，可以有效解决语音信号的特征提取和不等长语音匹配的问题。这些相关技术在孤立字识别领域有着不错的识别效果，取得了很大成功，从而掀起了语音识别研究的热潮。在 20 世纪 70 年代末期，Linda 等提出了矢量量化（Vector Quantization，VQ）[6]