语音识别：从入门到精通

课程大纲

1. 语音识别综述

1.1 什么是语音

1.2 语音产生与感知

1.3 语音处理技术的范畴

1.4 语音识别的定义

1.5 语音识别的历史

1.6 语音识别的应用

1.7 基础系统构成与关键技术

1.8 评价标准

1.9 课程介绍

语音识别是语音处理中的重要任务，也是人工智能应用热点。作为课程的开篇章节，这一部分将概述语音信号的相关基础知识、语音产生与感知的基本原理、语音识别技术的基本概念、发展历程、主流系统的基本构成与关键技术、评价标准等，同时对本课程的内容进行一个详细介绍。

2. 语音信号处理及特征提取

2.1 信号处理基础知识

2.2 语音特征提取

2.3 实战

一个语音识别系统的第一步，是对语音信号进行分析和特征提取。本节首先介绍语音信号时域和频域的基本特点，简要介绍频域分析方法（快速傅里叶变换），最后详细讲解当前语音识别系统广泛采用的特征，如 MFCC、FBANK 特征以及其提取的具体流程。

3. GMM 以及 EM 算法

3.1 GMM 模型

3.2 EM 算法

3.3 基于 EM 的 GMM 参数估计

3.4 实战

高斯混合模型（Gaussian mixture model，GMM）是在机器学习、图像识别、语音识别

等任务中被广泛采用的经典模型，在深度学习成功应用之前，基于 GMM 的概率分布模型一直用来建模声学特征的分布。本节课将详细介绍 GMM 模型原理和基于期望最大（Expectation maximization，EM）算法的参数估计过程，并结合 GMM 来深入探讨通用 EM 算法及其背后的机器学习原理。

4. HMM 模型

4.1. HMM 模型及其三个问题

4.2. GMM-HMM

4.3. 实战

隐马尔可夫模型(Hidden Markov Model, HMM)是可用于标注问题的统计学习模型。本节首先介绍隐马尔可夫模型的基本概念，然后分别介绍隐马尔可夫模型的概率计算算法、学习算法和预测算法，最后介绍 GMM-HMM 模型。

5. 基于 GMM-HMM 的语音识别系统

5.1 Big-picture（包括建模单元、三音素模型，训练，解码）

5.2 单音素模型训练（Baum-Welch 算法，Viterbi 算法训练，对齐）

5.3 三音素以及决策树

5.4 Viterbi 解码

5.5 实战

在本节课程中，我们会介绍特征提取、GMM 模型、HMM 模型在一个简单的孤立词识别系统语音中的基本应用。接着，我们引申到连续语音识别系统，分别引入 Pronunciation Model(Phone)、单音素模型、三音素模型、决策树等概念。与此同时，我们会深入探讨 HMM 的三个基本问题在语音识别系统中的实际实现和应用。

6. DNN-HMM 声学模型

6.1 DNN 简介

6.2 DNN-HMM 声学模型介绍

6.3 DNN-HMM 声学模型训练流程

6.4 基于各种 NN 的声学模型

6.5 实战

在传统语音识别 HMM-GMM 系统中，我们使用 GMM 对声学模型的概率密度分布进行建模。在这个深度神经网络 DNN 大爆发的时代，DNN 取代 GMM 成为更好的概率密度分布

的建模方法，从而过渡到目前主流的 HMM-DNN 语音识别系统。在本节课程中，我们引入 HMM-DNN 语音识别系统，介绍其基本的原理、神经网络的结构 (如 FNN、CNN、LSTM、TDNN 等)。

7. 语言模型

7.1 语言模型以及 n-gram 基础知识

7.2 N-gram 语言模型训练、回退等

7.3 RNN-LM

7.4 高级话题：大词汇量连续语音识别梳理

7.5 实战

在语音识别系统中，语言模型（language model）与声学模型具有同样重要的地位。正是由于基于统计的 N-gram 语言模型的产生，使语音识别任务从小规模的音素识别、指令识别、基于规则的识别时期迈向了大词汇量连续语音识别（large vocabulary continuous speech recognition, LVCSR)时代。统计语言模型的基本任务通过给定历史信息预测即将产生词的概率，N-gram 模型限定了历史信息的长度。如今，随着深度神经网络的发展，RNN，LSTM 等记忆性网络的提出，利用任意长历史信息变为可能，从而进一步帮助提升语言模型和语音识别任务的效果。本节课程讲详细介绍语言模型的基本原理，N-gram 语言模型的训练、多种回退算法、基于 RNN 的语言模型。

8. 基于 WFST 的解码器

8.5 WFST 的基本知识

8.6 WFST 的各种操作

8.7 基于 WFST 的解码器原理

8.8 高级话题：rescore

8.9 实战

解码器是语音识别任务中的最后一环，它担负着快速并准确的将声学模型（acoustic model，AM）和语言模型（language model，LM）结果有机结合，产生识别结果的重任。它的核心问题是如何产生准确的词网格（lattice），对一个高质量的词网格进行后处理可以进一步提升语音识别效果。此外，解码器也在基于嵌入式训练（Embedding training)的声学模型训练中起着至关重要的作用，将自动机理论引入解码器，使得解码器变得更加准确、清晰、易操作。本节课我们将介绍带权有限自动机(weighted finite-state transducer, WFST)的基本知识、基于 WFST 的解码器。

9. 区分性训练

9.1 区分性训练基本思想

9.2 区分性训练准则（MMI，bMMI，MPE，sMBR）

9.3 基于 GMM-HMM 的区分性训练

9.4 基于 DNN-HMM 的 Lattice-free MMI 训练

不同于最大似然准则聚焦于寻找最适合标注序列的模型，区分性训练旨寻找能将最优标注序列与其竞争序列相区分的模型。在传统的区分性训练中，需要解码生成 lattice 来近似所有与标注序列竞争的路径，在此过程中解码算法和反复生成 lattice 的计算开销很大。Lattice free MMI(LF-MMI)中通过构造 n-gram 的音素级语言模型来代替 lattice，并对模型拓扑结构和其他一些列的细节进行调整，从而有效的支撑 lattice-free 的区分性训练。目前，基于 LF-MMI 的训练已经在很多语音识别任务上取得 state-of-the-art 的效果。

10. 端到端语音识别

10.1 动机

10.2 Sequence-to-Sequence 与注意力机制

10.3 LAS

10.4 Speech Transformer

10.5 CTC

10.6 RNN-T

端到端语音识别(End-to-End ASR)是将传统语音识别多个组件整合进一个神经网络的语音识别技术。本节首先介绍端到端语音识别的动机，然后介绍 sequence-to-sequence 框架和注意力机制，最后介绍主流的端到端语音识别方法 LAS、Speech Transformer、CTC 和 RNN-T。

11. 总结展望

11.1 课程回顾

11.2 语音识别面临的挑战

11.3 语音识别前沿展望

作为本课程的最后章节，我们将回顾本门课程的知识点，对语音识别技术进行总结，并对未来语音识别技术的发展趋势进行展望。