一种新型英语口语自动打分系统

一．背景

随着计算机技术的发展，越来越多的学习软件可以帮助人们更方便地学习外语。目前绝大多数计算机辅助外语学习软件主要关注文字应用能力和语言理解能力的训练，却很少关注口语发音能力训练。应用语音处理技术，可以实现英语学习中的口语发音自动打分。

当前主流的英语口语打分系统分为整体打分系统和对比打分系统两种。整体打分系统不提供标准发音，直接测试发音人的发音标准程度，因而依赖一个背景标准发音模型；对比打分系统提供标准发音，发音人跟读标准发音，系统评价发音人发音与标准发音的相似程度。

本专利提出一种新的整体打分系统设计方法。整体打分系统一般包括声音信号的预处理和特征提取、基于标准发音库的统计模型建模、基于统计模型的发音对齐与概率计算等几个步骤，最后将所得概率归一成 0-1 之间的标准打分。传统整体打分方法一般基于高斯混模型（GMM）或隐马尔可夫-高斯混合模型（HMM-GMM）。

二．发明要点

本发明提出一种基于深度神经网络(DNN)后验概率特征的整体打分系统，基本思路是利用 DNN 的区分性建模特性，得到一种对噪声更加鲁棒的局部描述特征（帧后验概率），进而得到有效的句子全局特征，最后利用多层感知器（MLP）网络进行打分。

1. DNN 模型与 GMM 模型在英语口语打分中的区别

传统 GMM 或 HMM-GMM 模型的优化目标为对数据分布描述的精确性，即使得模型能最有效地解释数据的分布特性。DNN 是一种深度区分性神经网络模型，其优化目标为不同发音之间的区分性，即使得对不同发音的区分能力最大化。DNN 的这一特性使其可对抗背景噪声和信道影响。

另一个显著区别是，GMM 或 HMM-GMM 模型是统计模型，因此可以直接对句子进行建模和打分。相反，DNN 模型是一种非统计模型，不能直接对句子进行建模，只能对语音帧建模，得到局部特征，即帧后验概率。得到局部特征后，需要设计相应的全局特征提取函数，从局部特征中综合得到句子层的全局特征。最后，这些合局特征用来进行句子级的口语打分。

2. 基本流程

DNN 模型经过充分训练以后，给定一个语音特征向量帧 O(t)作为输入，其输出即为该语音帧对不同发音（包括噪音）的后验概率向量，记为 u(t)。在打分过程中，对各帧后验概率进行分布统计，提取全局特征，送入多层前向神经网络模型（MLP）进行区分性打分。

实际处理中，首先对待测语音 O 通过 DNN 提取到每帧后验概率{u(t)}，即基于 DNN 的局部特征向量。依{u(t)}，将语音 O 与音素 P 进行对齐，得到对齐结果 L(O,P)。依 L(O,P)得到每帧语音 O(t)在其对应的音素 P_t 上的后验概率，记为 $u(t,P_t)$。统计$\{u(t,P_t)\}$在 8 个取值区间的分布比例，形成一个 8 维的全局特征向量[s(1),s(2),…s(8)]，记为

$$s(i)=\frac{1}{T}\sum_{t}\delta(c(i-1)<u(t,P_t)\le c(i)) \qquad i=1,2,...8$$

其中 T 为语音 O 的总帧数，δ 为狄拉克函数，当参数中所设条件满足时取 1，否则取 0。{c(i); i=0,…8}是一个对概率取值区间[0,1]的划分。 考虑到 DNN 输出概率的非均匀性，我们取对数划分，即:

$$c(0)=0$$

$$c(i)=10^{8-i} \qquad i=1,2,3...,8$$

将特征向量[s(1),s(2),…s(i)]送入 MLP 模型，得到的输出即为对句子 O 的打分评价。

打分系统流程图如图（1）所示，其中所用到的 DNN 模型如图（2），MLP 模型如图(3)所示。

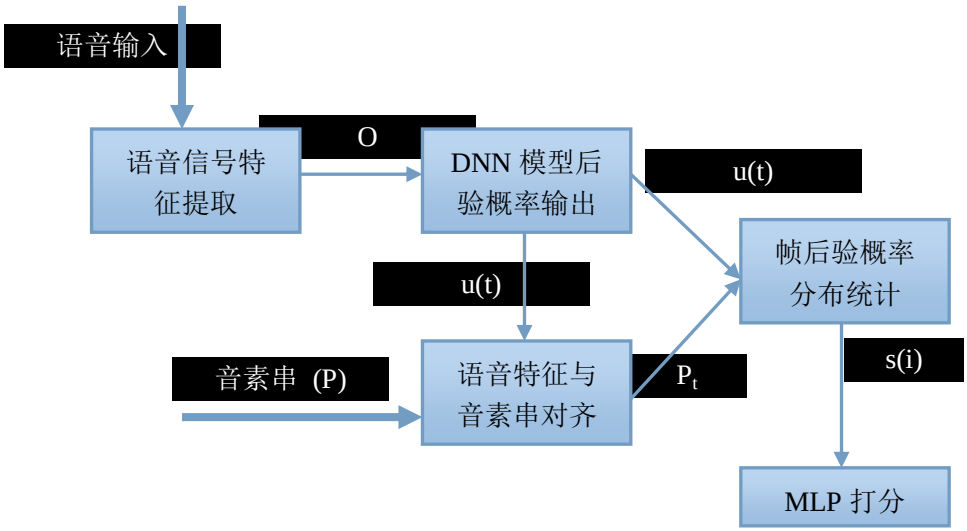

图 1：基于 DNN 的口语整体打分系统流程图

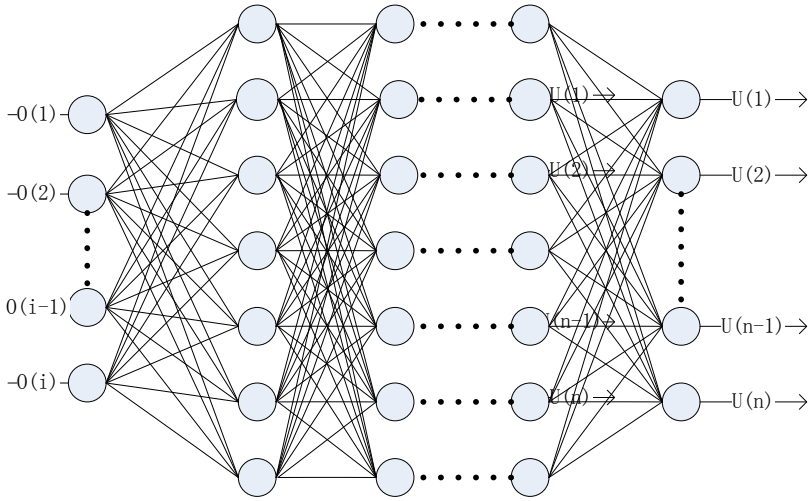

图 2: DNN 模型

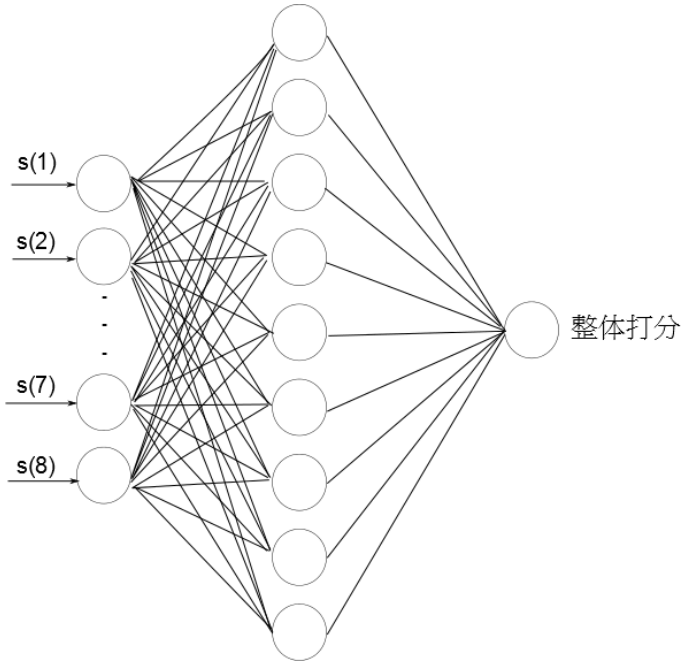

图 3: MLP 模型打分示意图

3. **模型训练**

   在训练 DNN 模型时，我们选择每帧语音对应的音素作为模型的目标输出，通过随机梯度下降算法进行模型参数优化。在训练 MLP 时，我们对训练集中的每个句子进行人工打分，以此打分作为 MLP 训练的目标输出。

三. 方案优势

1. DNN 是区分性模型，因此基于该模型的整体打分系统比传统基于 GMM 的打分系统具有更强的噪音和信道鲁棒性

2. MLP 是区分性模型，因而基于 MLP 的打分方法对发音质量亦具有更强的区分性。
3. MLP 基于人工标注的质量评价进行学习优化参数，因而得到的分数分布更加合理。