英语自动打分研究

Automatic Scoring for English Utterances (CSLT-TRP-140001)

胡博（Bo Hu）

2014/11/20

CSLT, RIIT, Tsinghua Univ.

Pachira Ltd.

1. 背景简介

随着计算机技术的发展，计算机辅助教学已成为现代教育技术在教育领域运用的一个重要方面。越来越多的学习软件已经在帮助人们学习外语。计算机丰富的图形、图像、声音处理功能有力的促进了人们的语言学习效果。但是目前很多计算机辅助语言学习软件都主要关注语言的文字应用能力和语音理解能力的训练，却很少关注语音的口头表达能力训练。口语学习主要表现在发音的学习上。而语音技术的不断成熟则为辅助学习者发音提供可能，广阔的语言学习市场正吸引着众多的开发商投入到语音技术/语音识别技术研制语言教学产品中来。

本研究关注对英语发音学习过程中的口语发音评分技术。当前主流的口语打分系统分为通用打分系统和对比打分系统两种。通用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，发音人一标准发音进行模仿，系统评价发音人发音与标准发音的相似程度。

通用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，最后将所得概率归一成 0-1 之间的标准打分。传统通用打分系统一般基于 HMM-GMM 模型，分数归一可以使用 sigmoid 函数。

对比打分系统一般包括模板收集，语音信号预处理与特征提

取，基于模板的信号对齐等三个模块。对齐方法一般采用动态规划算法，如 DTW。近来通常采用基于 HMM-MAP 的统计模型方法对模板建立简单的 HMM 模型，再通过计算测试语音在该模型上的概率值得到对比分数。本质上，通用打分和对比打分并没有明显分界，对齐打分可以看做是通用打分在训练语料极度稀缺条件下的近似。然而在实际应用中，因为对齐打分可供对比的模板有限，难以生成复杂的概率模型，因而一般采用近似的模板直接匹配(DTW)或模型自适应(HMM-MAP)方法。

本文对口语评测中的通用打分和对比打分技术进行细致的研究。首先我们研究基于深度神经网络的通用打分技术。通过用大规模标准发音的预料库训练的深度神经网络，我们得到每一语音帧的后验概率向量，近而形成整句发音的后验概率谱。通过计算正确发音音素串在该后验概率谱上的概率值分布特征，得到精确的通用打分。

同时，我们研究适用于跟读的对比打分技术。传统对比打分使用 MFCC 等标准语音特征进行匹配度计算。由于 MFCC 特征包含信道、噪音、口音、情绪等多种因素，得到的打分不仅受噪音的影响显著，且很难对发音的语调、韵律、音色等进行精细刻画。本文研究将发音的音调、节奏、音色进行区分性打分的技术，同时研究噪音鲁棒性特征的提取方法，对比不同特征的贡献。

2. 英语口语评价的通用打分技术

通用打分依拉于一个标准发音模型。传统发音模型多基于HMM-GMM 框架，属于产生式模型，容易受噪音影响。通用打分的目的在于计算某一发音对标准发音的距离，使其距离正确发音尽可能近，而距离错误发音尽可能远，因而本质上是对不同发音进行区分，因此更适合用区分性模型进行建模。本研究中，我们采用 DNN 模型进行声学建模，该模型的输入为一个语音帧，输出为该语音帧对不同发音（包括噪音）的相似程度（后验概率）。在打分过程中，语音信号与标准发音的音素通过 Viterbi 算法对齐，即可计算出该语音与 DNN 模型的匹配程度。

实际处理中，单纯 DNN 的得分是所有语音帧的平均分值，难以衡量每个单词的具体匹配程度。同时，对语音帧进行简单平均会丢失大量匹配信息，因此我们对 DNN 对各帧的打分计算各种不同的统计量，特别是计算每个音素的平均分值和这些分值的分布数据，形成多维特征，并将这些特征进行区分性建模，得到最后的打分。本文选择 MLP 作为区分性建模方法。

2.1 DNN 模型训练

我们选择 wsj 数据库进行标准英文发音建模。该数据库含有约 100 小时的英文朗读语料，男女各约 10 人。声音采样由麦克风

完成，采样率为 16k 赫兹，采样精度为 16 bit.

基于 kaldi 工具，我们首先建立一个基于 MFCC 的 GMM 模型，进而训练一个输入为 40 维 FBank 特征，输出为 4000 个音素状态的 DNN 模型。模型训练采用 SGD 算法，价差熵作为训练目标函数。

2.2 通用打分模型训练

打分模型目的在于对实际场景中的输入语音进行打分，因此需要训练一个区分性模型来对 DNN 的输出进行分值计算。如前所述，我们计算 DNN 的分值统计量，形成高维特征向量。我们采用区间分布概率作为主要的特征。具体而言，我们将语音帧和音素对齐，并依据对齐结果得到每个语音帧的 DNN 分值。将每个语音帧的 DNN 分值归入均分的 8 个概率值区间，计算每个区间分布的语音帧比例，得到 8 维特征。这些特征作为输入，将该句的认为评定打分作为输出，即得到 MLP 的一个训练样本。

本文中，我们采用 5 个档次的分值标注(1-5)，每一分值收集约 1500 句训练语音进行人工标注，依上所述构造训练样本，训练得到 MLP 模型作为打分模型。

上述方法可以很容易扩展到单词打分，即对单词内的语音帧分布

进行统计，依次对整句打分进行调整得到单词打分。

3. 英语口语评价的对比打分技术

对比打分计算待测试句子与标准发音句子的相似程度，并归一化为打分值。由于句子相似度的概念相对模糊，我们将打分分为三个方面：节奏打分，音调打分与音色打分。

我们首先采集了一个训练集和测试集。先选定一个文本集，对其中每个文本，每个说话人朗读至少 2 遍，将同一个人的同一句子为相似，不同人的相同句子标为不相似。

3.1 节奏打分

节奏描述发音过程中的发音长度变化，即每个音素发音长短的相关性。我们采用线性模型来描述两句话之间的节奏关系。

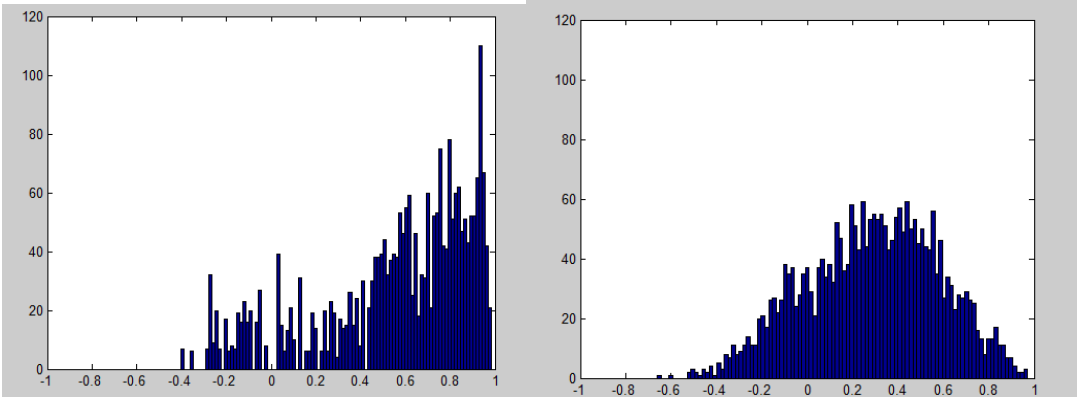
首先对待测语音和标准语音利用通用打分技术进行音素对齐，得到每个音素的时间长度。设句子中共含有 N 个音素（注意待测语音和标准语音的音素个数相等），将每句话的音素长度作为一个随机变量，计算待测语音音素长度与标准语音音素长度的线性相关性。

设标准语音音素长度变量 x，待测语音音素长度变量为 y，并假设两者符合线性关系:

$$y = a x + b$$

因为句子中有 N 个音素，我们可以通过 N 个(x,y)的采样点来估计上面线性模型中的 a,b。同时，我们可以计算 y 与 x 的相关系数 c，对应于上面线性模型的残差。显然，a,b,c 这三个变量与 x,y 的相似度直接相关，可以用来计算语音的节奏得分。

图 (1) 给出了基于相关系数 c 的分布图。选取测试语音与标准语音为同一人进行相关系数计算，给出的相关系数分布如左图所示;选取测试语音与标准语音为不同人进行相关系数计算，给出的相关系数分布如右图所示。可以看到，相同说话人得到相关系数统计上明显高于不同说话人，证明相关系数在区分说话人方面具有显著性。当考虑 a,b,c 三个特征向量的时候，我们采用 MLP 作为归一化模型，将三个特征向量归一化为 0-1 之间的打分。



相同人相同文本语音　　不同人相同文本语音

图 (1). 基于节奏的同一说话人与不同说话人的相关系数分布

3.2 音调打分

音调描述发音的基频随时间变化的规律。我们采用同节奏打分相似的方法，提取不同音素的基频，并计算测试语音与标准语音的线性相关性模型，提取特征参数 a,b,c 进行 MLP 建模。

考虑到发音的变化以及辅音的基频不清晰情况，我们在线性模型建模中不计入辅音。根据通用打分技术生产的音素对齐结果，将每一个元音音素平均分为 3 段，分别计算这 3 段语音的基频平均值，从而得到整句话的基频向量。

图（2）给出相同说话人与不同说话人基于基频的相关系数分布。我们看到基频的相关系数对不同说话人具有明显的区分性。同样，当考虑 a,b,c 三个特征向量时，我们采用 MLP 模型来对打分进行归一。

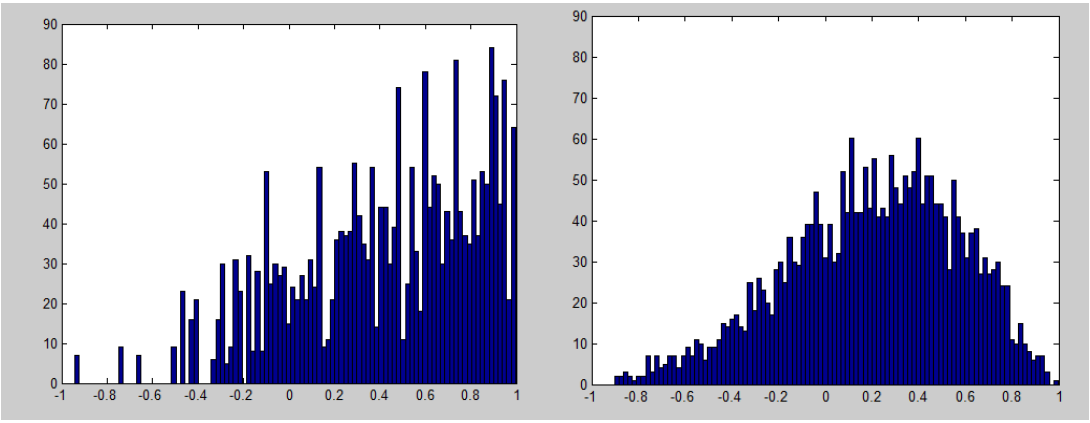


图 (2). 基于基频的同一说话人与不同说话人的相关系数分布

3.3 音色打分

音色是描述说话人发音特质的变量。音色具有模糊性，广义的音色包括节奏和音调，狭义的音色与声音中共振峰的分布相关性更强。特别是，我们认为音色更关注静态相似性，这点与节奏和音调更强调变量的变化相关性有所差别。因此，我们选择静态模型来对音色进行打分。

我们的音色打分分为两种方法：第一种方法借鉴说话人识别方法，用 GMM 模型进行打分。这一方法忽略语序(GMM 模型不考虑语音帧的前后顺序)，因此不需要音素对齐。由于模型采用 MFCC 基础特征，受信道和噪声影响较大。第二种方法显式提取基频和共振峰位置和强度信息，将其在两个发音序列上的差异作为特征向量进行 MLP 建模，从而预测音色差异。由于基频和共振峰等和说话人相关的信息被显式提取出来，这一方法较基于 MFCC 的 GMM 模型具有更强的抗噪声能力。另一方面，由于将基频和共振峰的差异作为特征，因而该方法和发音内容及顺序相关，需要进行音素对齐。

3.3.1 基于 MFCC+GMM 的音色打分方法

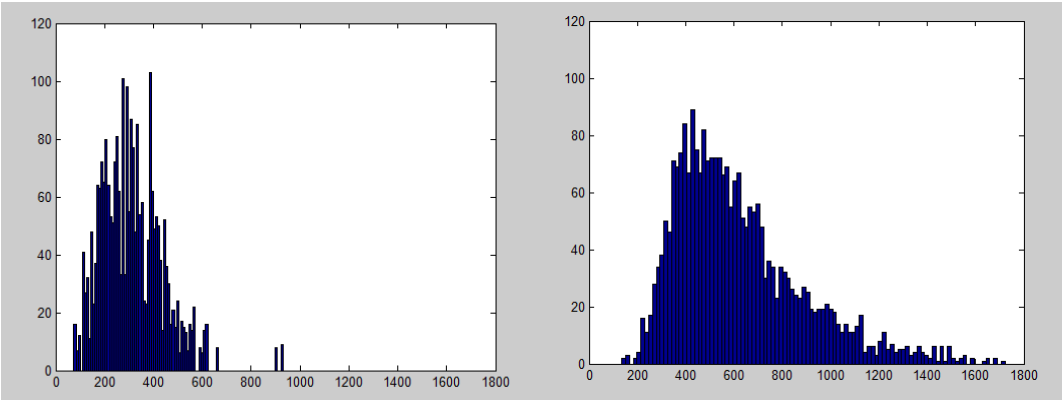
我们首先基于一个较大规模的语料库(wsj)训练得到一个通用混

合高斯模型 UBM。对标准语音，通过 MAP 方法对 UBM 进行训练，得到标准说话人模型 GMM。对待测语音，将其分别在 UBM 和 GMM 上计算概率，最后计算两者的比值，比值越大说明越接近。通过一个 sigmoid 函数，可以将该比值归一化为音色打分。

3.3.2 基于共振峰的音色打分方法

将语音通过频谱分析计算每一帧语音的基频和共振峰，得到 10 维共振峰值。根据通用打分技术得到的音素对齐结果，将共振峰值以音素为单位进行时间平均，得到整句话的音素共振峰分布。将对比打分的两句话对应的元音音素求共振峰向量的距离，进而平均得到整句话的共振峰距离。注意，因辅音音素的共振峰不明显，故在计算距离时不予考虑。

根据距离大小可以判定两句话的音色相似度，距离越小音色越相近。根据相关性分析，频谱的前 4 个峰值的频谱距离相关性最大，所以实验中选择前 4 维峰值作为计算距离的向量。图（3）给出相同人和不同人的语音共振峰距离分布。可以看到，相同人的共振峰距离要小于不同人的共振峰距离，分布具有明显区分性。



相同人相同文本语音　　不同人相同文本的语音

图 (3). 基于共振峰距离的同一说话人与不同说话人的相关系数分布

4. 总结

本文详述了基于 DNN 和区分性建模的英文口语打分技术，特别是提出基于 DNN 帧分值统计特性的打分技术和基于 MLP 的特征融合技术。实验表明该方法可以有效提高打分性能，并可提供多角度的打分值。

附件 A：

A.1 代码库：

git clone git@192.168.0.51:speech/pingce.git

A.2 通用打分方法

a) 安装：cd pingce/linux; ./install.sh

b) 生成可持行文件:　cd linux_rhythm;

cp ../pachira_score/lib/libpachira_score.so .; make

c) 打分: ./score_stream <scp file>，结果显示到单词长度及打分，以及整句话打分。

scp file 文件格式为:

sample.wav| Today I want to tell you three stories from my life |sample.fea

其中第一个|之前为音频，两个|之间为音频对应的文本，第二个|之后为对齐特征文件（注意：第一个|之前不能有空格，第二个|之后不能有空格），为了计算音调，音色，节奏使用。

A.3 对比打分方法

A3.1 音调，节奏和共振峰音色打分方法

a) cd pingce/compare_demo; ./install.sh

b) cd demo; ./run.sh

demo.sh 中接收两个对比音频文件及其相应的音素对齐文件。

打分显示为

```
harmonic_score = 98.40
pitch_score = 56.77
rhythm_phone_score = 90.22
rhythm_word_score = 95.37
```

pitch_score 为音调打分

rhythm_phone_score 为以音素为单位的节奏打分

rhythm_word_score 为以单词为单位的节奏打分

harmonic_score 为基于共振峰的音色打分

A.3.2 GMM-MAP 的音色打分方法

a) cd pingce/compare_demo/gmm; make，

b) ./gmm-score model/final.dubm ../1.wav ../2.wav。