第五届全国人机语音通讯学术会议(NCMMSC'98)，286-289，98年7月26~31日，哈尔滨

本文获NCMMSC'98优秀论文二等奖

汉语语音水平评价方法研究

徐明星 宋战江 郑方 吴文虎

(北京 清华大学计算机科学与技术系100084)

摘 要：汉语语音水平评价是汉语语音识别的一个分支，它具有广泛的应用前景。本文以CDCPM模型为基础，针对模型的特性，提出了两种语音水平评价方法——直接评价方法和间接评价方法，并把它们综合应用于一个实际的汉语语音水平评价系统之中，取得了很好的实验结果。结果表明，本文提出的语音水平评价方法是可行的和有效的。

关键词：语音水平评价，CDCPM模型，CDN分布

一、引言

汉语语音水平评价是汉语语音识别的一个分支，其主要目的是利用计算机对测试人员输入的汉语语音的规范程度给出一个比较客观的评价。这可以分为指定文本的评价和不定文本的评价两个大类。由于中国的人口众多，地域宽广，有许多不同的方言，而且各地的汉语语音普通话教学的水平也存在着很大的差异，这种现状显然很不利于普通话的推广，不利于人与人之间的语言交流；另外，鉴于中国的国际地位日益提高，中国市场潜力巨大，越来越多的外国人希望掌握汉语的发音技巧，因此，依据汉语语音识别的基本原理，利用计算机来帮助校正/提高汉语规范程度，具有十分重要的现实意义和广泛的应用前景。

本文对汉语语音水平评价方法进行了初步探索和研究。本文组织结构如下：第二部分给出了直接评价方法；第三部分则提出了间接评价方法，第四部分针对两种方法的不同特性，提出了综合利用的形式和方法；最后给出了评价系统的测试结果。

二、直接评价方法

1、基本思路

在对汉语语音规范程度进行评价的时候，一种自然的想法是：将待评价的原始输入语音同标准的语音模型进行直接匹配比较，根据它们的吻合程度，得到对待测语音的评价得分。由于这种评价方法是直接进行的，只涉及到原始输入语音所对应的标准模型，所以我们称之为直接评价方法。该方法大致的工作流程示意图如下所示：

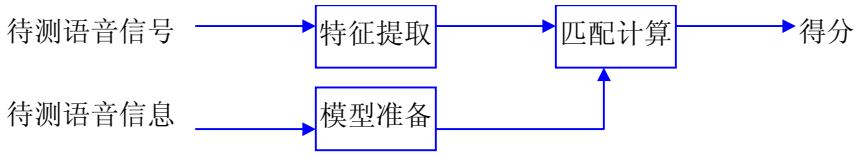

图1 直接评价方法工作流程示意图

其中，“待测语音信号”是指用户按照要求输入的语音采样数据，“待测语音信息”是指应用指定的输入语音的汉语音节信息，“特征提取”的功能是根据模型的要求，对输入的原始语音进行特征提取计算，如计算信号的能量、过零率、基音周期、LPC-CEP[1]等参数，一般在特征提取的时候需对原始语音进行分帧处理。“模型准备”的功能是依据待测原始语音的音节信息，将需要的音节模型参数提取出来，为下一步的匹配计算做好准备。“匹配计算”的功能是按照模型的特点，对待测的原始语音的特征参数进行处理，得到输入语音同规范语音之间的匹配程度，从而得到评价分数。

2、标准模型的建立

语音标准模型的建立方法很多，现在最为常见的是基于HMM[1]思想的语音模型，如DHMM[2]、CHMM[3]、SCHMM[4]等，还有一些是标准HMM的修正改进型，也可以拿来使用。在本文建立的系统中，标准语音模型所采用的是CDCPM[5]模型，它是HMM的一种简化模型，没

有状态转移矩阵A，因为状态的转移是从左至右直接进行的，而且在状态发生转移时，状态不允许进行跨越转移。为描述CDCPM模型的各个状态，我们引入了若干子状态，以EMM[6]（嵌入式多模板）的方式组织起来，各个子状态则用一个CDN分布[6]来表征，这些CDN分布的参数就组成了整个语音模型的模型参数。

3、匹配计算

对CDCPM模型中的“中心距离正态分布”(CDN)

$$N_{\mathrm{CD}}(x;\mu,D)=\frac{2}{\pi D}\mathrm{e}^{-(x-\mu)^2/\pi D^2},x\in[0,\infty) \tag{1}$$

来讲（μ和D是CDN分布的参数），根据概率论中关于置信区间的理论[7]，由$\sigma=\sqrt{2}\pi D/2\approx 1.25D$，可以得到，在区域$[0,2.5D]$中分布有95%的样本。如果我们将区间表达式中的“2.5”设置成一个阈值k来控制区间的大小，则可以控制落入区间[0, kD]中的样本数目。从特征样本与该区间的关系，我们可以得到语音样本同标准的关系。

由于输入语音进行了分帧处理，各帧计算相应的特征参数，为了表述方便，设$\boldsymbol{O}=\{\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T\}$表示原始输入语音的特征序列，其中$T$是原始语音发音的时间长度（以帧数计）。设$\Lambda=\{\mu_{nm},D_{nm}|1\le n\le N,1\le m\le M\}$表示语音模型的参数，其中$N$为模型的状态数，$M$为描述模型状态中所含的子状态的数目。

在进行匹配得分计算之前，先要划分特征序列$\boldsymbol{O}$中各个特征向量o_t的状态归属。划分状态的方法很多，本文采用的基于等特征变化量原则的“非线性分段”[8]法，它简单、高效。由于CDCPM模型对状态序列的特殊要求，$\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T$将被归入到N个状态中，其状态序号的排列将是单调非减的序列。

A、方法一：定义发音序列第t帧特征矢量$\boldsymbol{o}_t$对模型Λ的评价分数为：

$$S(\boldsymbol{o}_t|\Lambda)=\begin{cases}1, & \text{当}\max\limits_{1\le m\le M}\{d(\boldsymbol{o}_t,\mu_{nm}|\boldsymbol{o}_t\in f\ (n,m))\}\in[0,kD_{nm}]\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中，$f(n,m)$表示模型Λ第n个状态中的第m个子状态，$d(\cdot,\cdot)$是特征矢量的距离度量函数，k是置信区域大小控制参数。于是，特征矢量序列$\boldsymbol{O}$对模型Λ的评价分数为：

$$S(\mathbf{O}|\Lambda)=\sum_{t=1}^{T}S(\mathbf{o}_t|\Lambda)/T \tag{3}$$

B、方法二：如果考虑状态中所有的子状态，则特征矢量$\boldsymbol{o}_t$对模型Λ的评价分数可定义为：

$$S(\boldsymbol{o}_t|\Lambda)=\begin{cases}1, & \text{当}\sum\limits_{m=1}^{M}S(\boldsymbol{o}_t|f(n,m))>TH\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中，$S(\boldsymbol{o}_t|f(n,m))=\begin{cases}1, & \text{当}d(\boldsymbol{o}_t,\mu_{nm}|\boldsymbol{o}_t\in f\ (n,m))\in[0,kD_{nm}]\\ 0, & \text{其他}\end{cases}$ (5)

TH是一个在1到$M-1$之间变化的整数阈值，一般由经验给出。

C、方法三：在实际系统中，各状态可能要用不同数目的子状态来描述[9]，而方法一和方法二都不能反映这个特性。为此，我们重新定义特征矢量序列$\boldsymbol{O}$对模型Λ的评价分数为：

$$S(\mathbf{O}|\Lambda)=\sum_{t=1}^{T}\sum_{m=1}^{M(n(t))}S(\mathbf{o}_t|f(n(t),m))/\sum_{t=1}^{T}M(n(t)) \tag{6}$$

其中，$n(t)$表示第t帧特征矢量所属的状态，$M(n(t))$表示状态$n(t)$中的子状态的数目。

三、间接评价方法

1、基本思路

前面所述的直接评价方法，直接利用待测原始语音和标准语音模型的匹配分数，考虑的是单一的对应程度，即输入的原始语音的特征只与相应的标准模型进行比较。如果我们同时也考察一下原始输入语音特征同其他标准模型的关系，由于这些匹配分数实际上就是原始语音对各个标准模型的识别分数，反映的是原始语音对标准模型的匹配程度，所以，对这些信息进行映射转换处理，同样也能得到反映原始语音标准程度的评价分数。

2、计算过程

设一共有K个语音标准模型，$\Lambda_1,\Lambda_2,\cdots,\Lambda_K$ 代表这K个标准模型，$RS_1,RS_2,\cdots,RS_K$ 分别表示原始语音的特征序列 $\boldsymbol{O}$ 对这K个标准模型的识别分数（显然，如果是语音识别系统，则按照要求给出前几个得分最高的模型作为识别候选就可以了），则对输入语音的规范程度评价分数为：

$$Score(\boldsymbol{O}) = F(RS_1,RS_2,\cdots,RS_K,k) \tag{7}$$

其中，F(...)表示映射变换函数，k是输入语音所对应的标准模型的序号。F变换依据标准模型识别分数相互之间的关系，得到关于原始输入语音的评价分数。因为 RS_k 代表了原始语音的特征序列 $\boldsymbol{O}$ 对模型 Λ_k 的匹配程度，所以可以设计如下的映射变换函数F:

$$F(RS_1,RS_2,\cdots,RS_K,k) = \big(K - Order(RS_k)\big)/K \tag{8}$$

上式中，函数 $Order(RS_k)$ 表示第k个模型识别分数（RS_k）在全部识别分数中的大小排名，函数的值域为[0, K)。

需要说明的是，RS_k 的计算同前面所论述的直接评价方法中的计算方法是不同的。在直接评价方法中，利用的是概率论置信区间的理论，而 RS_k 的计算则不考虑这些，它利用的是CDN分布的概率计算公式[5]，即：

$$RS_k = \prod_{t=1}^{T} \max_{1\le m\le M}(N_{CD}^{k}(o_t;\mu_{n(t)m},D_{n(t)m})) \tag{9}$$

其中，n(t)表示第t帧特征 o_t 所归属的状态序号，N_{CD}^{k} 是模型 Λ_k 的CDN分布的密度函数。

四、进一步讨论

在对汉语语音评价系统的测试实验中，我们发现，两种方法都有评价不够稳定的时候。显然，直接评价方法和间接评价方法是两种不同的评价体系，它们依据的原理各不相同，计算方法更是差别很大，因而可以考虑把它们综合起来，以便提高评价系统的稳定性；另外，由于这两种方法一个强调对自身的匹配程度，一个强调相互之间的关系，有一定的互补性，从这一点出发，它们也应该结合在一起。

在对两种评价方法进行综合的时候，需要考虑它们给出的评价分数之间的关系。从前面的讨论和相应的计算公式可以看出，这两种方法得到的评价分数的变化特性是一致的，即输入的语音越是规范，评价得分就越高，因此综合的时候可以采用加法或乘法来实现。在本文所作的实验测试中，使用乘法系统的性能更稳定些。即：

$$Score(\boldsymbol{O}) = Score1(\boldsymbol{O}) * Score2(\boldsymbol{O}) \tag{10}$$

其中，$Score1(\boldsymbol{O})$ 表示使用直接评价方法所得到的评价分数，$Score2(\boldsymbol{O})$ 表示使用间接评价方法所得到的评价分数。

汉语是一种有调语言，所以在语音水平测试系统中，需要考虑说话人的音调问题。从语音中提取音调的方法很多，我们采用了小波变换的方法提取音调[10]。但是，由于变调现象在汉语语言交流中比较普遍，不能把音调错误当作严重错误来处理，所以，我们的系统如果检测到音调错误，只是酌情扣除一些分数，并留有一个控制开关来设置音调评测的严格程度，这样就比较实用了。另外，评测系统的稳定性在一定程度上取决于对语音音节切分的正确性，我们在系统中引入了一些控制切分的新方法和新规则，详情将另文发表。

五、结语

在本文的实验中，测试了300个汉语句子，覆盖了汉语全部单音节。语音数据的采样率为16K，语音帧的大小为32ms，帧移为16ms。系统的评价性能是用测试人员的主观感受来统计得出的。参加评测的人员一共有20人，来自不同的地域，有的还是外国人。根据测试人员使用后的评测意见（对方言口音和外国口音的评价性能），系统的性能如下：

表1 系统性能统计结果

系统性能	很好	好	良	差
比例	5%	45%	35%	15%

上述结果表明：尽管测试工作还很不严格和周密，但按照本文提出的评价方法，系统对输入汉语语音规范程度的评价能力还是得到了用户的认可的。

参 考 文 献

1、 杨行峻、迟惠生等　语音信号数字处理. 1995，电子工业出版社.

2、 L.R.Rabiner, S.E.Levinson, and M.M Sondhi, “On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition,” Bell Syst.Tech.J., vol.62, pp.1075-1105

3、 L.R.Bahl, P.F.Brown, P.V.de Souza, and K.L.Mercer, “Speech Recognition with Continuous-Parameter Hidden Markov Models,” Readings in Speech Recognition, pp.332-339, edited by Alex Waibel & Kai-Fu Lee, 1990.

4、 X.D.Huang & M.A.Jack, “Semi-Continuous Hidden Markov Models for Speech signals,” Computer Speech and Language(1989), 3:239-251

5、 郑方，吴文虎，方棣棠.　CDCPM及其在语音识别中的应用. 软件学报，1996, 7: 69~75

6、 郑方. 连续无限制语音流中关键词识别方法研究：[博士学位论文]. 北京：清华大学计算机系，1997

7、 盛骤、谢式千、潘承毅等 概率论与数理统计 (第2版) 1988，高等教育出版社.

8、 蒋力 基于概率统计模型的非特定人语音识别方法与系统的研究：[硕士学位论文]. 北京，清华大学计算机系，1989

9、 Fang Zheng, Mingxing Xu, Wenhu Wu, “The Description of the Intra-State Feature Space in Speech Recognition,” Int’l Conf. Research on Computational Linguistics, pp.272-276, Aug.22-24, 1997, Taiwan

10、 黄仁中 基音检测和汉语四声判别：[硕士学位论文]. 北京，清华大学计算机系，1996