一种新的汉语连续语音声调评测算法

沈彩凤，俞一彪
(苏州大学电子信息学院语音技术研究室，江苏苏州 215006)

摘要：提出一种新的连续语音的声调评测算法，该算法可应用于计算机辅助语言学习系统和普通话水平测试中的声调评测。考虑到连续语音声调受上下文之间的相互影响，采用三音节单元建立高斯混合模型(Gaussian Mixture Model, GMM)，三音节中辅音部分用 Spline 插值法拟合声调曲线来反映音节间基音频率的转移信息，并利用 Fujisaki 模型去除语句的语调和说话人个性特征，只对基频曲线中的声调特征建模。实验结果显示，相比于传统方法，采用三音节 Spline 插值和 Fujisaki 改进特征的方法使得机器与人工打分的相似度在测试集中分别提高了 8.75%和 14.09%。

关键字：声调评测；连续语音；Spline 插值；Fujisaki 模型；高斯混合模型
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1000-3630(2013)-04-0305-07
DOI 编码：10.3969/j.issn1000-3630.2013.04.010

A novel tone evaluation algorithm for Chinese continuous speech

SHEN Cai-feng, YU Yi-biao
(*Speech Technology Laboratory*, *Soochow University*, *Suzhou* 215006, *Jiangsu*, *China*)

Abstract: A new algorithm of objective tone evaluation for Chinese mandarin continuous speech is proposed, which can be used for the tone pronunciation training in Computer Assisted Language Learning (CALL) system and the test of Chinese mandarin speech named as Putonghua Shuiping Ceshi (PSC). A syllable's tone is influenced by context in continuous speech. Therefore, it is reasonable to use tri-syllables as basic units to train GMM (Gaussian Mixture Model) of tones. To get the transition information from the previous voiced region to the current one or from the current to the next voiced region, the pitch value of unvoiced region is interpolated with Spline function. Based on the Fujisaki model, only the lexical tone from the F0 contour is extracted to train GMM. The experimental results show that the correlations between subject and object evaluations based on Spline interpolation and Fujisaki model are improved by 8.75% and 14.09% respectively, comparing to the traditional features.

Key words: tone evaluation; continuous speech; Spline interpolation; Fujisaki-model; Gaussian Mixture Model

0 引 言

声母、韵母和声调是汉语音节的三要素，因此人们不仅仅凭借着声韵母来确定字词的意义，还凭借不同的音节声调来分辨它们，如果不考虑声调对音节的影响，汉语中 1300 多带调的音节可以减少成 400 多无调的音节。因此，在计算机辅助语言学习(Computer Assisted Language Learning, CALL)以及普通话水平测试(Putonghua Shuiping Ceshi, PSC)中，不仅强调对汉语音节的发音的正确性，也越来越多地关注普通话声调的错误与否。

汉语音节的声调特征在孤立发音时，通常都是比较稳定，按照五度值[1]描述系统可表达为：阴平(55)，阳平(35)，上声(214)，去声(51)，但是这种标准声调值也只出现在孤立字发音情况下。在几个音节组成的短语或者在连续语音发音中，各音节原来的声调特征因为受到了前后音节的影响使得轮廓曲线发生变化，这种情况下声调 F0 轨迹曲线比孤立音节中的形状和特征复杂得多[2]。2007 年，汤霖等人在分析汉语声调特点的基础上，提出消除语速和音节间影响的建模方法，选择反映声调特点的五个基音频率比值和归一化基音频率一起作为声调评测特征建立高斯模型，结果客观评测与主观打分的符合率达到了 88.24%[3]。2008 年，潘逸倩等人提出了基于韵律信息的连续语流调型评测研究，以韵律词为基本建模单元，建立基于多空间概率分布的 HMM 调型模型 MSD-HMM，针对有河南与山东方言背景的非标准发音，机器评分与专家评分相关度达到 0.661 和 0.695[4]。

考虑到连续语音发音中前后音节对当前音节

收稿日期：2012-05-10；修回日期：2012-08-27
基金项目：国家自然科学基金资助项目(61271360)；苏州市应用基础研究计划资助项目(SYG201230)
作者简介：沈彩凤(1986－)，女，江苏宿迁人，硕士研究生，研究方向为汉语语音评测。
通迅作者：沈彩凤, E-mail: mantianxing45610@126.com

声调的影响，本文以三音节为声调单元建立 GMM 模型，对测试语音计算其后验概率并通过映射给出声调的客观评分。

1 声调评测系统概述

本文的声调评测系统如图 1 所示。

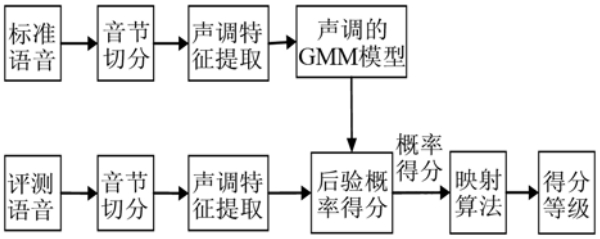

图 1 声调评测流程框图
Fig.1 Flow chart of tone evaluation

声调评测中，首先对连续语音进行切分获得每一个音节在语音中的起止点。然后以音节为单元提取语音的基音频率轨迹，并从得到的连续三音节基音频率轨迹曲线中提取声调特征并进行高斯混合模型(GMM)的训练。评测语音经过相同的预处理过程得到三音节声调特征，最后根据三音节 GMM 模型计算声调的后验概率，并按照一定规则进一步映射为具体的评价得分输出。

音节切分采用 HMM 的方法来获得音节的边界。HMM 切分(Forced Alignment, FA)是根据给定的声学 HMM 模型，按照维特比算法(Viterbi Algorithm)通过调整各音节 HMM 的边界以达到似然度最大化来实现。HMM 模型由大量的连续语音语料以汉语声韵母为基元，采用 13 维 MFCC 及其一阶 MFCC 差分和二阶 MFCC 差分构成 39 维特征向量训练得到。

2 特征提取及其改进方法

2.1 轮廓分段特征

在进行声调特征提取时，必须考虑不同说话人的基音频率动态范围影响，因此，为了消除说话人声道特性，本文在特征提取之前先对说话人的基音频率进行归一化，比例归一化公式如(1)。

$$f_{\mathrm{nor}}=\frac{(Max-Min)\cdot f+Min\cdot f_{\max}-Max\cdot f_{\min}}{f_{\max}-f_{\min}} \tag{1}$$

式(1)中：Max 和 Min 为归一化需要达到的最大和最小基音频率值，$f_{\max}$ 和 $f_{\min}$ 是语音中的最高和最低基音频率值，f 为当前基音频率，f_{nor} 为归一化之后的基音频率。

文献[5]中描述了忽略声调的细节特征对声调识别有很大的提高，而声调的主要值和声调趋势是声调识别的重要因素，因为这些声调的轮廓特征可以减少连续语句中协同发音、音节重音及其句子语调的影响。本文基于三音节的声调轮廓特征提取方法是将每个单音节浊音段的 F0 曲线平均分为相同长度的几个子段，第 k 个子段的基音向量为 $\{F0_{ks},F0_{ks+1},\cdots,F0_{ke}\}$，对于这个子段内的特征提取可以有两种方法，分别是：子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征和子段斜率、截距的特征。这两种特征提取的方法描述[6,7]如下。

(a) 子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征

基音频率的一阶差分向量为 $\{\Delta F0_{ks},\Delta F0_{ks+1},\cdots,\Delta F0_{ke}\}$，其中 ks 和 ke 是当前子段的起始和结束帧。该子段内的特征如表达式(2)、(3)所示。

$$\overline{F0}_k=\frac{1}{ke-ks+1}\sum_{i=ks}^{ke}F0_i \tag{2}$$

$$\Delta\overline{F0}_k=\frac{1}{ke-ks+1}\sum_{i=ks}^{ke}\Delta F0_i \tag{3}$$

第 k 个子段的特征值是该子段内的基音平均值 $\overline{F0}_k$ 及其一阶差分平均值 $\Delta\overline{F0}_k$，如果音节被分成了 M 个子段，则 $\{\overline{F0}_1,\Delta\overline{F0}_1,\overline{F0}_2,\Delta\overline{F0}_2,\cdots,\overline{F0}_M,\Delta\overline{F0}_M,\Delta\overline{F0}\}$ 为音节 $2M+1$ 个整体特征向量，其中 $\Delta\overline{F0}$ 是整个音节声调的一阶基音差分平均值。

(b) 子段斜率、截距的特征

用线性回归的直线拟合第 k 个字段的 $F0$ 曲线，即是用 $f(n)=an+b$ 的直线去逼近第 k 个字段的 $F0$ 曲线，其中 a 是直线的斜率，b 是截距。假设 $X=[0,1,\cdots,ke-ks]$ 以及 $Y=[F0_{ks},F0_{ks+1},\cdots,F0_{ke}]$，那么：

$$(a,b)=\arg\min_{a,b}\sum_{n=ks}^{ke}(aX_{n-ks}+b-Y_n) \tag{4}$$

如果音节被分成了 M 个子段，那么这个音节的 $2M$ 个整体特征向量为 $\{a_1,b_1,a_2,b_2,...,a_M,b_M\}$。

2.2 特征的插值改进

在单音节中，$F0$ 曲线是对音节中的浊音部分进行基音检测得到的，基音频率可以作为单字声调的全部特征。然而在连续语音中，每个音节的声调都受到上下文的影响，传统 F0 曲线的求取方法也只得到独立音节的频率，却忽略两个音节之间基音频率的转换信息。以音节为单位的基音频率可以得到浊辅音的基音值，所以如果当前音节的韵母是浊辅音，那么它与前一个音节间的频率是连续的，这两个音节间的转移信息也是可以描述的。但如果当前音节的韵母为清辅音，那么它与前一个音节间的频率就会出现间隔，这样也就丢失了两个音节之间的

频率转移特征[8]。

本文在音节基音提取的基础上，以三音节声调作为一个处理单元，对每两个音节间 *F*0 为 0 的帧进行插值处理，以得到详细的声调变化特征。图 2 给出了对语音“一种树”提取的 *F*0 并进行三种方法插值的比较结果图(见图 2)，这三种方法分别是：Spline 三次样条插值、线性插值和分段三次 Hermite 插值。

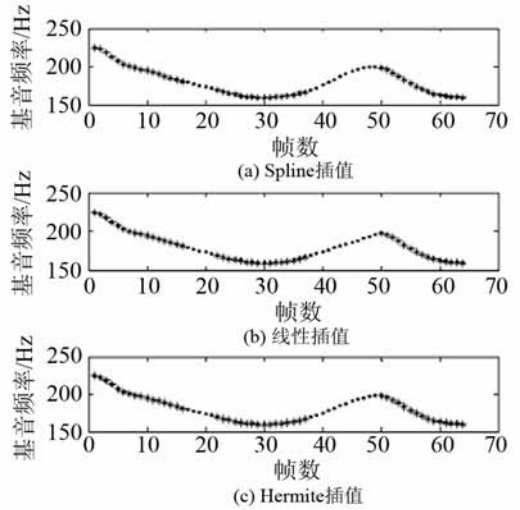

图 2　“一种树”三音节基音插值图

Fig.2　Interpolation of a tri-syllable example

图 2 中粗黑星号标志点为程序提取的每帧对应的基音频率，细小的圆点是在已有声调基础上对基音频率为 0 的帧进行插值得到的。从图 2 的三个波形比较中，本文选择运用 Spline 插值(图 2(a))的方法，Spline 插值的曲线光滑度比线性插值(图 2(b))的曲线有很大提高，同时比 Hermite 插值(图 2(c))方法简单易于实现。

2.3　声调特征的 Fujisaki 模型改进

连续语流的 *F*0 曲线的轮廓含有很多超音段的信息，因此 *F*0 曲线的变化不仅体现了音节声调的特征，也体现了连续语音流中韵律变化的趋势。也就是说 *F*0 曲线同时携带了语音的声调信息和语句的语调信息。根据赵元任先生的说法，连续语音语调与单字声调的关系就好比波浪和涟漪一样是层层叠加的关系，可以用代数总和表示[1]。

说话是在呼气阶段利用声门下肺气的压力作原动力来进行发音的。在其他的生理条件如声带长短与厚薄相同的前提下，说话人声门下压力在开始时比较大，因此音高也较高，越往后，声门下的压力就会越小，音高也就很难保持在起始时的高度。除非换一口气再重新发音，或者尽力去维持或提高声门气压。由此可知，语句总体的音阶运动走势在这种生理因素下是逐渐下倾的，并且这种走势在发音的换气或节奏的转变时，才会由于音阶的跃变重新再来。Fujisak[9,10]提出以生理学为基础利用喉部结构及它的相互作用来描述 F0 生成和控制的声调模型，这个模型利用重叠组织方法较好地描绘了语句的这种下倾走势。

Fujisaki 模型将基频曲线拆解成三个不同的元件函数的叠加，并分别找到相对应的发声器官的物理特性来解释这些元件函数。这三个元件函数分别为短语元件(Phrase Component)、强调元件(Accent Component)和基底频率(*F*b)。其中短语元件反应较大单位基频曲线的控制发声限制，即语句的语调信息；*F*b 是基本音高，代表说话人的个性信息；强调元件反应较小单位基频曲线的控制发声限制，即是每个音节的声调信息。Fujisaki 模型从生理上、声学特性上以及韵律控制上对语调做出了清楚的描述，这种描述也符合赵元任所说的大波浪和小涟漪的关系。

本文提出利用 Fujisaki 模型将语句的共同特征语气和说话人个别特征 *F*b 剔除掉，只对语句中的音节声调建模。采用 Mixdorff 提出的方法[11]来解析原来的 *F*0 曲线，方法如下：

(1) 提取语句的 *F*0 曲线，对曲线中清音和无声段进行 Spline 插值，并对特征取对数。

(2) 以一组截止频率为 0.5Hz 的高通滤波器来分离 *F*0 曲线，自动提取出 *F*0 曲线中变化比较大的部分即为音节的音调信息。输出定义为高通曲线(HFC)。

(3) 扣掉高通部分剩余的平滑曲线则定义为低通曲线(LFC)，此部分曲线变化和缓，为语音语调和说话人个性信息的和。

图 3 中是利用 Fujisaki 模型将音节声调与语句语调和 *F*b 分离的结果，图 3(a)的虚线为基频插值之后的 *F*0 轨迹，实线表示分离之后的低通曲线，

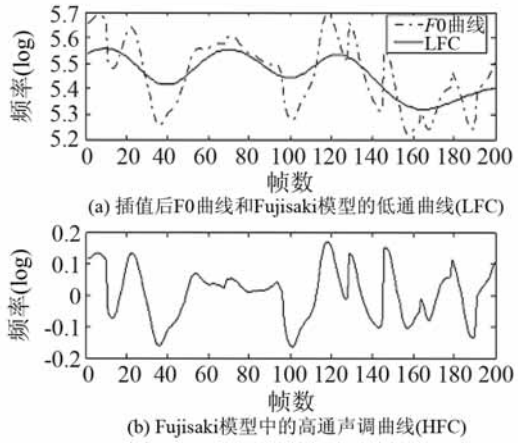

图 3　Fujisaki 模型的声调提取

Fig.3　Tone extraction based on Fujisaki model

即语句的语调和说话人特征 *F*b 的叠加，图 3(b)的波形是 Fujisaki 模型的高通部分，也就是音节的声调曲线。

3 声调的 GMM 模型

汉语音声调包括轻声可以分为 5 种，对这 5 种声调采用 GMM 模型来描述它们的分布。因为模型是以三音节为基础建立的，因此在句子的句首和句尾分别插入无声段，则 GMM 模型总数为 180。

M 阶混合高斯模型的概率密度是由 *M* 个正态分布的高斯概率密度函数的线性组合表示的，如式(5)所示：

$$P(X/\lambda)=\sum_{i=1}^{M} w_i N(\boldsymbol{X},\mu_i,\Sigma_i) \tag{5}$$

式(5)中 $P(\boldsymbol{X}/\lambda)$ 是声调特征向量 $\boldsymbol{X}$ 在相应的声调模型 λ 下的概率密度函数，输出概率密度函数中的 w_i 表示第 i 个混合高斯函数的权，满足 $\sum_{i=1}^{M} w_i=1$，$N(\boldsymbol{X},\mu_i,\boldsymbol{\Sigma}_i)$ 是第 i 个混合高斯的概率密度函数，其中 μ_i 代表均值向量，$\boldsymbol{\Sigma}_i$ 是协方差矩阵，如式(6)：

$$N(\boldsymbol{X},\mu_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu_i)^t\boldsymbol{\Sigma}_i^{-1}(X-\mu_i)\right\} \tag{6}$$

$\boldsymbol{X}$ 为输入的声调特征矢量，本文 GMM 模型对应的特征矢量分别为传统轮廓子段特征、Spline 插值改进特征和 Fujisaki 改进特征，完整的混合高斯模型由上述参数组合起来表示为 $\lambda=\{\omega_i,\mu_i,\boldsymbol{\Sigma}_i\}$，$i=1,\cdots,M$。GMM 模型的训练就是给定一组训练数据，依据 EM(期望最大化)准则确定模型 λ 的参数[12]。

4 声调评测算法

声调评测基于声调 GMM 模型[13]进行，使用后验概率(Posterior Probability, PP)作为声调的似然度得分，指定声调 *ref* 相对声调特征 $f0$ 的输出后验概率为 $p(ref|f0)$，根据贝叶斯公式，它与声调特征 $f0$ 对于 *ref* 声调 GMM 模型的输出似然度 $p(f0|ref)$ 有关系，如公式(7)所示[14]。

$$p(ref|f0)=\frac{p(f0|ref)p(ref)}{p(f0)}=\frac{p(f0|ref)p(ref)}{\sum_{k=1}^{M}p(f0|t_k)p(t_k)}\approx\frac{p(f0|ref)}{\sum_{k=1}^{M}p(f0|t_k)} \tag{7}$$

式(7)中，$\sum_{k=1}^{M}p(f0|t_k)$ 是 *M* 个声调特征 $f0$ 的 GMM 模型的观察概率总和，并且式中最后一个等号是在满足 $p(ref)=p(t_k)$ 的条件下成立的，即当文本中 5 种声调出现的概率相同时。本文中 $p(t_k)$ 则由参与训练的各声调数目统计得到，这将在第 5 节中加以介绍。

以上得到的概率值并不适合作为系统评分直接反馈给用户，因为它显然与人类听觉主观评价有差异，所以有必要将概率值映射为人们所能接受的评测打分。本文针对不同发音人所发的声调特征，将概率得分映射为三个等级，即声调正确、声调缺陷和声调错误。

最简单的映射策略是根据上述得到的后验概率计算对数似然度并求绝对值 *PP*，设置全局门限进行等级的评定，即对于指定声调 *ref* 的声调特征等级划定符合式(8)。

$$PP(ref)\begin{cases}<Thresh1 & \text{发音正确}\\ >Thresh2 & \text{发音错误}\\ \text{others} & \text{发音缺陷}\end{cases} \tag{8}$$

式(8)虽然简单，但是实际应用中两个阈值的划分对声调模型精度的依赖性比较大，*PP* 分布图如图 4 所示。

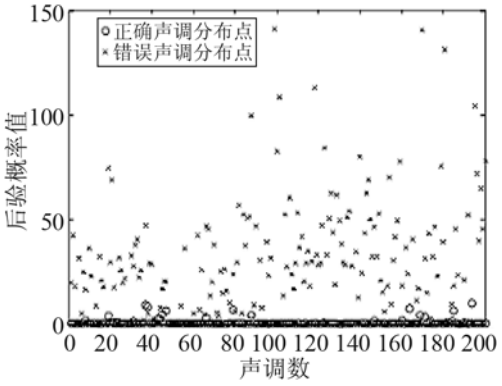

图 4 声调 *PP* 分布图

Fig.4 Distribution map of tone posterior probabilities

图 4 中圆圈形表示对应指定声调 *ref* 的 *PP*，叉形表示对于 *ref* 剩下的其他声调的 *PP* 中的最小值，也就是识别中最接近 *ref* 的错误声调。因为声调识别模型精度问题，很明显图 4 中圈与叉在很大程度上是很难分辨的。为提高判断精度，也为了更好地找出判断阈值。本文针对发音正确的声调，统计其一维高斯分布，即对于 *N* 个声调正确音节的 *PP* 得分 $s_1,s_2,\cdots,s_N$ 组成训练样本 $correct\_s=(s_1,s_2,\cdots,s_N)$，计算它们一维高斯概率分布 $p(s/correct\_s)$，用这个概率分布作为评价的依据。对任意声调得分 ms_i，它的评测映射规则为：

如果 $|\log(p(ms_i/correct\_s))|<\alpha_1$，则认为该音节声调发音正确。

如果 $|\log(p(ms_i/correct\_s))|>\alpha_2$，则认为该音节声调发音错误。

否则，认为该音节是发音缺陷。

判定阈值 α_1 是由正确声调的 *PP* 对于高斯分布统计得到的，而 α_2 是上述错误声调的 *PP* 对于高斯分布统计得到的，它们对于模型的分布如图 5 所示。

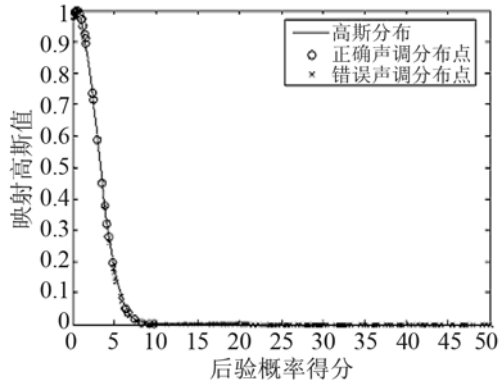

图 5　声调 *PP* 高斯分布图

Fig.5　Gaussian distribution map of tone posterior probabilities

图 5 中圈与叉号标志与图 4 中描述是相同的，曲线则表示正确音节 *PP* 得分的一维高斯分布图。图中可以看出，对于 *PP* 得分从 10 开始，正确与错误声调的分布有较为明显的分界。

5　实验及其结果

5.1　实验数据库

本实验中所用语音均是在实验室的录音室中采用 DegiDesign C|24 专业设备进行录制，语音噪声低，纯净度较高，语音采样率为 16k，量化位数 16bit。标准语音是录自 10 个(5 男 5 女)普通话水平比较好的同学，朗读文本选自普通话水平测试文本中的 2 篇文章，包括 960 个音节，其中声调阴平 172 个，阳平 239 个，上声 146 个，去声 307 个，轻声 93 个。根据朗读的文档中声调的分布状况，第 5 节中 5 种声调 T0、T1、T2、T3、T4 的先验概率分别为 0.098、0.179、0.249、0.153、0.321。评测系统的性能用系统打分与人工打分的相关度来评价。打分语音录自 3 个(2 个女生 1 个男生)普通话发音水平一般的同学，语音的声调标准程度参差不齐。为了比较算法的性能，本文首先对这些打分语音中每个音节的声调进行正确、缺陷、错误这三个等级的人工评测，再计算这些人工打分和算法中机器打分的相关度。相关度算法如式(9)所示。

$$\overline{A}=\frac{1}{N}\sum_{i=1}^{N}A_i\ ,\quad \overline{B}=\frac{1}{N}\sum_{i=1}^{N}B_i$$

$$Corr_{A,B}=\frac{\sum_{i=1}^{N}(A_i-\overline{A})(B_i-\overline{B})}{\sqrt{\sum_{i=1}^{N}(A_i-\overline{A})^2}\sqrt{\sum_{i=1}^{N}(B_i-\overline{B})^2}} \tag{9}$$

式(9)中，A_1、A_2、…、A_N；B_1、B_2、…、B_N 分别是 N 个测试者的人工和机器的打分。$Corr_{A,B}$ 为它们之间的相关度。

本实验的基本模型是三音节声调曲线分段轮廓特征的 GMM 模型，在此模型的基础上，验证第 2 节中两种改进的声调特征，目标是分析算法的机器打分与人工打分的相关性是否得到提高。

5.2　实验结果及其分析

因为每个人的语音是已经录制好的，为了消除不同发音人声道的影响，首先运用公式(1)将每个发音人的声调基音频率归一化到 150~250Hz 之间。用 2.1 节中子段特征提取算法提取传统的轮廓子段特征。本文中，对子段 $\overline{F0}$，$\Delta\overline{F0}$ 的特征提取时，当前音节被分为 4 段得到其 9 阶特征矢量，前一个和后一个音节均被分成 3 段得到它们单独的 7 阶特征矢量，最后用得到的全部 23 阶特征矢量训练三音节的 GMM 模型。而子段斜率、截距的特征因为比子段 $\overline{F0}$，$\overline{F0}$ 的特征少了单个音节的整体特征 $\Delta\overline{F0}$，所以相同的分段方法得到的特征矢量也少了 3 个，这样三音节的 GMM 模型的特征矢量维数是 20。由该 GMM 模型的后验概率作为其概率得分，最后通过映射算法，并根据第 5.1 节系统性能评价方法，计算算法得分等级和人工打分等级的相关性。机器打分和人工打分的相似度越大，说明机器分的准确性越高。在这个实验基础上，针对不同特征的改进进行讨论。

实验一：根据 2.2 节声调特征的插值改进，将 Spline 插值算法运用于 F0 曲线转移信息的获取以改进声调特征，并与上面的子段 $\overline{F0}$，$\Delta\overline{F0}$ 的特征和子段斜率、截距的特征两种特征模型方法进行相关性比较，结果如表 1 所示。

表 1　插值特征与传统特征比较

Table 1　Correlation difference between interpolation and traditional features

特征	相关度	
	训练集	测试集
传统子段 $\overline{F0}$，$\Delta\overline{F0}$ 的特征	0.7392	0.6542
插值子段 $\overline{F0}$，$\Delta\overline{F0}$ 的特征	0.8021	0.6643
传统子段斜率、截距的特征	0.7105	0.6721
插值子段斜率、截距的特征	0.7424	0.7309

从表 1 中可以看出，包含声调转移特征的插值方法不管是对于子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征还是子段斜率、截距的特征在训练和测试集中相关性都有所提高，并且在子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征的相关度在训练集中尤其明显，从 0.7392 提高到了 0.8021。而在子段斜率和截距特征对于训练和测试集的提高都很

大。这说明插值的频率点作为声调转移部分的特征能很好地描述三音节声调中连音对前后音节声调的影响，从而更精确地描述基音轨迹变化曲线，对评测结果有很大的提高。

实验二：根据 2.3 节声调特征的 Fujisaki 模型改进中的方法对传统子段特征进行改进。首先对提取的 F0 轨迹曲线进行 Fujisaki 模型的拆解，提取只包含音节声调信息的高通曲线 HFC。然后再进行子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征和子段斜率、截距的特征的提取。该实验中因为说话人个性信息包含在已经丢弃的低通曲线 LFC 中，所以不需要进行说话人声道归一化处理。该方法与传统子段特征进行比较的结果如表 2 所示。

表 2 Fujisaki 特征与传统特征比较

Table 2 Correlation difference between Fujisaki and traditional features

特征	相关度	
	训练集	测试集
传统子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征	0.7392	0.6542
Fujisaki 子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征	0.7903	0.7464
传统子段斜率、截距的特征	0.7105	0.6721
Fujisaki 子段斜率、截距的特征	0.8024	0.7361

从表 2 中可以看出，在利用 Fujisaki 模型去除语句语调特征和说话人个性特征这些干扰特征，仅仅对于音节的声调特征进行评测时，相比于传统子段的特征，不管是训练集还是测试集，人工打分和机器打分的相关度都有很大程度的提高。尤其是 Fujisaki 子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征在打分测试集中提高了 0.09。并且对于 Fujisaki 的模型，不管采用哪种子段特征的方法，训练集和测试集数据的相关度都差不多。这说明，基音轨迹中包含的语句语调特征和说话人特征很大程度影响了声调评测的结果，而 Fujisaki 模型能很好的提取出音节的声调特征，剔除了语句的语调和说话人信息。

表 1 和表 2 是对改进后的特征与传统特征相比较的表，比较这两张表可以发现：在训练集上，Fujisak 子段 $\overline{F0}$，$\Delta\overline{F0}$ 的特征比插值子段 $\overline{F0}$、$\Delta\overline{F0}$ 的特征有 0.01 的降低，这是因为 $\overline{F0}$、$\Delta\overline{F0}$ 的特征能更好地描述变化剧烈的 $F0$ 曲线，而 Fujisaki 模型去除了语调特征，使得 $F0$ 曲线平滑性有很大改善。为了比较这三种特征对相关度的影响，图 6 给出了三种特征在测试集上的的比较柱形图。

图 6 中，左边的三个柱形表示的是子段 $\overline{F0}$、$\Delta\overline{F0}$ 及其改进后特征在打分测试集上的相关度，右边三个柱形表示的是子段斜率、截距及其改进后特征在打分测试集上的相关度。图中黑色表示的是传统子段特征，灰色表示的是 Spline 插值改进的子段特征，白色表示的是 Fujisaki 模型下子段的特征。从柱形图看出 Spline 插值的方法简单容易，对段斜率、截距的特征有明显提高，但是对子段 $\overline{F0}$、$\Delta\overline{F0}$ 特征的改进不明显，而 Fujisaki 模型虽然复杂，但是对两种特征都能达到很好的效果，并且使得子段 $\overline{F0}$、$\Delta\overline{F0}$ 特征的相关度达到最大。

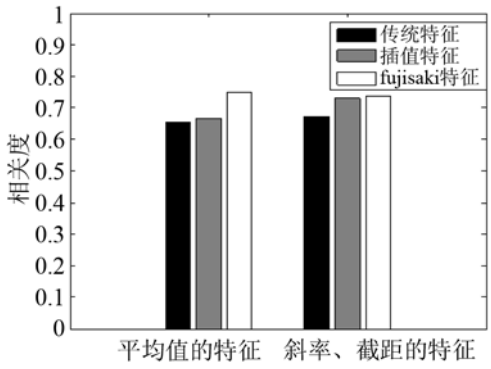

图 6 三种特征的相关度比较图

Fig.6 Correlation diagram of the three characteristics

6 结 论

以孤立音节求取 F0 曲线并以三音节为训练特征单元的 GMM 模型能很好地给出语音的声调得分。实验证明，运用 Spline 插值对语音辅音段进行基音补充的方法，在测试集中打分相关度从 0.6721 提高到 0.7309。同时从利用 Fujisaki 模型只提取语句的音节声调曲线，以减少语调特性和说话人个性特征的实验结果中可以看出，这种改进方法在测试集中比传统方法增加了 0.09 的打分相关度，提高了声调评测的准确性，相信随着训练语料的增加会使得打分结果得到很大改善。

参 考 文 献

[1] Yuen Renchao. A grammar of spoken Chinese[M]. Univ of California Pr, 1986:16-39.

[2] 王安红. 汉语声调特征教学探讨[J]. 语言教学与研究, 2006, **6**(3): 70-75.
WANG Anhong. On teaching chinese tone feature[J]. Language Teaching and Linguistic Studies, 2006, **6**(3): 70-75.

[3] 汤霖, 尹俊勋. 普通话声调的客观评测[J]. 中文信息学报, 2007, **21**(6): 116-124.
TANG Lin, YIN Junxun. Objective evaluation of putonghua tones[J]. Journal of Chinese Information Processing, 2007, **21**(6): 116-124.

[4] 潘逸倩, 魏思, 王仁华. 基于韵律信息的连续语流调型评测研究[J]. 中文信息学报, 2008, **22**(4): 88-93.
PAN Yiqian, WEI Si, Wang Renhua. Tone evaluation of Chinese continuous speech based on prosodic words[J]. Journal of Chinese Information Processing, 2008, **22**(4): 88-93.

[5] YE Tian, ZHOU Jianlai, CHU Min. Eric Chang. Tone recognition with fractionized models and outlined features[C]// Proc. of ICASSP, 2004: 105-108.
[6] PAN Fuping, ZHAO Qingwei, YAN Yonghong. Improvements in tone pronunciation scoring for strongly accented mandarin speech[C]// International Symposium on Chinese Spoken Language Processing, 2006.
[7] ZHANG Junbo, WU Hemin, YAN Yonghong. Tone Pronunciation Quality Scoring of Mandarin Multi-syllable Words[C]// ICSP, 2010: 545-548.
[8] CHEN Jiangcun. A study on pronunciation assessment and tone recognition in mandarin Chinese[D]. Tsinghua University, 2008.
[9] Fujisaki H. Hirose K. Analysis of voice fundamental frequency Contours for declarative sentences of Japanese[J]. J. Acoust. Soc. Jpn., 1984, **5**(4): 233-241.
[10] Fujisaki H, WANG Changfu, Sumio Ohno, GU Wentao. Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model[J]. Speeeh Communication, 2005, **47**(1-2): 59-70.
[11] Hansjorg Mixdorff, Hiroya Fujisaki, et al. Towards the Automatic Extraction of Fujisaki model Parameters for Mandarin[C]// EuroSpeech, 2003: 873-876.
[12] Jeff A, Bilmes A. Gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models[C]// International Computer Science Institute, 1998.
[13] CHEN Jiangchun, Jyh-Shing Roger Jang, Jun-Yi Li, et al. Automatic Pronunciation Assessment for Mandarin Chinese[C]// IEEE International Conference on Multimedia and Expo, 2004: 1979-1982.
[14] Neumeyer L, Franco H, Digalakis V, et al. Automatic scoring pronunciation quality[C]// Speech Communication 30, 2000: 83-93.

• 简 讯 •

湖北恩施土家族苗族自治州文化中心大剧院竣工开演

湖北恩施文化中心是现代都市院章奎生声学设计研究所在我国华中地区承接建声设计的又一项演艺建筑工程项目，也是恩施土家族苗族自治州的标志性文化建筑工程项目。剧场容座为 1300 座，设有楼座和两侧跌落包厢，剧院有多用途使用要求，可以满足歌剧、戏剧、舞蹈、交响乐及会议等使用功能。2013 年 8 月 19 日恰逢自治州成立 30 周年，剧场作为 30 年庆的主会场具有重要的政治意义，经两年时间的设计、施工及调试于 2013 年上半年已基本建成并试演出。7 月下旬应业主邀请，声学所一行六人由项目负责人宋拥民博士带领并安排携带许多贵重声学测试仪器乘动车前往当地现场。章奎生教授也应业主总指挥向伟先生热情邀请首次亲赴恩施现场，参与音质检测和音质评价工作。为使这次检测评价工作更具客观性和可靠性，声学所还特地由业主专门邀请上影集团的著名录音师任大铭先生和上海歌舞团的音响师史汇荣先生同去现场参与剧场主观听音评价，两位都是国家一级录音师，在业内有“金耳朵”之称，另外还请了两位电声专业人士，协助电声系统检测及评价事宜。

在恩施文化中心剧场音质检测工作中分别对大剧场的空场和满场条件；对舞台上无乐罩和安装音乐反射罩条件；对戏剧和音乐会两种不同演出条件等都安排进行了音质测量，其中还专门安排了对扩声系统性能指标的检测工作。7 月 28、29 日两天的检测工作进行得十分全面、系统、详细，28 日夜间测量工作一直持续到凌晨四点。28 日晚上业主特地安排了主观听音评价测试音乐会，请当地青年民族乐团演出音乐会，演出后业主还召开了音质评价座谈会，建声、电声及录音师专家、演奏人员及乐团指挥等均参与了主观评价座谈，并提交了书面音质评价意见表，评价会一直开到半夜 12 点才结束。

通过现场音质测量和音质评价讨论，一致认为恩施文化中心剧场具有优良的音质效果，全厅平均混响时间达 1.6s，声音响度足够，声场均匀扩散、混响特性优良，厅内具有足够的前次反射声和侧向反射声，听音既有足够的清晰度也有必要的丰满度，且厅内十分安静，无噪声及振动声干扰，完全达到了建声设计预期的指标要求，表明建声设计取得了圆满成功，业主对本工程的建声设计及音质效果也表示了十分赞赏和满意。

这次声学所赴恩施现场检测和评价工作，业主非常重视，十分热情，向总还亲自安排食宿和交通，自治州的秦副州长及工程部相关领导都会见了章奎生所长，当地电视台也专门在现场电视采访了章奎生教授，扩大了华东建筑设计研究院有限公司章奎生声学设计研究所在湖北当地和土家族苗族自治州的影响，并已有新的建声设计工程项目准备再次委托声学所承担。

华东建筑设计研究院有限公司章奎生声学设计研究所　章奎生