第 35 卷 第 8 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.8 Computer Engineering April 2009
·博士论文· 文章编号：1000—3428(2009)08—0026—02 文献标识码：A 中图分类号：TN912.34

基于听觉模型的特征在英语重音检测中的应用

陈 楠，贺前华，李 韬
(华南理工大学电子与信息学院，广州 510641)

摘 要：对于英语等“重音节拍语言”，重音是一个非常重要的韵律学特征。从听觉模型的角度出发，利用基音同步幅度峰值特征能同时表征瞬时频率和强度信息的特点进行重音检测。使用基音同步幅度峰值特征以及与传统特征的组合对英语连续语音的试验结果表明，新特征能使系统误识率降低 1.5%。

关键词：重音检测；听觉模型；基音同步幅度峰值

Application of Auditory Model-based Feature in English Lexical Stress Detection

CHEN Nan, HE Qian-hua, LI Tao
(School of Electronic and Information, South China University of Technology, Guangzhou 510641)

【Abstract】 Stress is an important prosodic feature for stress-timed language such as English. This paper presents a new approach using auditory model-based Pitch Synchronous Peak Amplitude(PSPA) feature, which incorporates frequency and intensity information in English lexical stress detection. PSPA feature, along with traditional features and their combinations to English lexical stress detection are evaluated with ISLE database. Experimental results show that the combination of new feature and traditional features demonstrates a 1.5% error rate reduction.

【Key words】 stress detection; auditory model; Pitch Synchronous Peak Amplitude(PSPA)

1 概述

英语是一种“重音节拍语言”(stress-timed language)，重音既是英语语音结构的组成部分，又具有区别词义和词性的功能，同时还是语调和说话节奏结构的基础，因此，重音在英语中起着极为重要的作用。

目前，学术界对重音的定义和分类还存在争议，根据文献[1]的定义，英语重音是“比其他音节或单词更重要”的音节或单词。文献[2]将重音划分为音节层面的词重音和句子层面的节奏重音与语义重音。本文关注的是英语音节层面的重音检测，所指的重音即英语词 重音。

由于重音是一个在语音上难以确定的特征，因此它没有一个属于自己的语音特性。一般，重音是音高、音长、音强等特征的综合体，各特征对重音检测的贡献度有主次之分。文献[3]通过决策树等方法，验证了各特征对英语词重音知觉的贡献从大到小依次为音长、音强和音高。在实际应用中，通常使用时长、能量和基音分别表征音长、音强和音高特征。

各国学者对重音检测的研究始于 20 世纪 50 年代，在经历了 50 多年的研究后，重音检测实现了从孤立词(双音节单词对)到连续语音的发展。根据现有技术，在连续语音中对重音的识别率已达到 80%左右，但这一结果尚未达到令人满意的程度。由于重音检测中常用特征除了表征重音信息外，还携带多种其他语音信息，影响了重音检测的效果，因此寻找能更精确表征重音的特征是提高重音识别率的重要途径。

近年来，基于听觉模型的语音特征在语音信号处理各领域得到广泛应用，这是因为听觉模型模拟了人耳对声音信号的处理过程，提取的特征能反映听觉系统的独特性质。本文从模拟人耳听觉模型的角度出发，将基于听觉模型的基音同步幅度峰值特征引入重音检测中，提高现有系统的性能。

2 基音同步幅度峰值特征

文献[5]在文献[4]的 ZCPA 听觉模型基础上，提出了基于听觉模型的基音同步幅度峰值特征，并将其应用于抗噪语音识别中，该特征提取流程如图 1 所示。

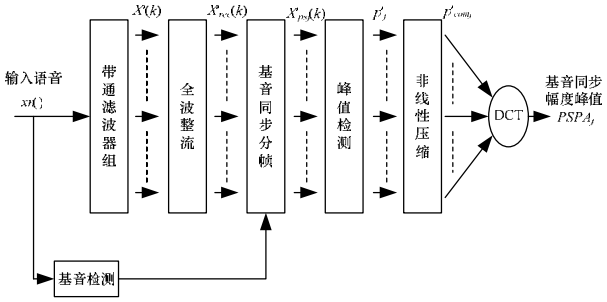

图 1 基音同步幅度峰值特征提取流程

该系统由基音检测、带通滤波器组、全波整流、基音同步分帧、峰值检测和非线性压缩等部分组成。20 路线性相位 FIR 滤波器组成的带通滤波器组用于仿真耳蜗基底膜对语音信号的时频分析特性；全波整流、基音同步分帧、峰值检测和非线性压缩用于仿真听觉神经纤维的功能。语音信号经过带通滤波器组后分为 20 路信号，分别进行全波整流和基音同步分帧，然后通过峰值检测和非线性压缩将信号的强度信息

基金项目：国家自然科学基金资助项目(60572141, 60602014)
作者简介：陈 楠(1981－)，男，博士研究生，主研方向：语音质量客观评价，重音检测；贺前华，教授、博士生导师；李 韬，助教
收稿日期：2008-09-20 **E-mail：**chn2000@163.com

和频率信息合成，最后将 20 路信号组合构成基音同步幅度峰值特征矢量。

为了利用听觉神经系统的基音同步机制，首先对滤波器组输出的 20 路信号进行基音同步动态分帧处理。根据实时检测的基音信息对语音进行清/浊分类。对于清音段，令帧长为 10 ms，帧移为 5 ms；对于浊音段，令帧长为当前基音周期的 2 倍，帧移为当前基音周期。

为了仿真听觉神经纤维的锁相程度和激励源密度之间的关系，在基音同步幅度峰值特征的提取中常用一个单调递增函数表示这种关系，即：

$$f(x)=\mathrm{lb}(1+x) \tag{1}$$

其中，x 表示帧内最大峰值。实际上式(1)实现了振幅的非线性压缩，模拟人耳的非线性处理过程。第 j 帧第 i 路经过非线性压缩的幅度峰值 $p_{com\,j}^{\ i}$ 的数学表达式如下：

$$p_{com\,j}^{\ i}=\left(\mathrm{lb}\left(1+p_{j,1}^{i}\right)+\mathrm{lb}\left(1+p_{j,2}^{i}\right)\right)\Big/2 \tag{2}$$

其中，

$$p_{j,1}^{i}=\max x_{ps\,j}^{\ i}(k)\quad 0\le k\le\frac{l_j}{2}+\delta \tag{3}$$

$$p_{j,2}^{i}=\max x_{ps\,j}^{\ i}(k)\quad \frac{l_j}{2}-\delta\le k\le l_j \tag{4}$$

l_j 为第 j 帧的帧长。对于浊音段，帧长为 2 倍基音周期，$p_{j,1}^{i}$ 和 $p_{j,2}^{i}$ 分别表示各基音周期内的信号峰值，δ 为微调值，一般为 2 ms～3 ms。

在进行非线性压缩后，将 20 路信号组合构成基音同步幅度峰值特征矢量 ***PSPA***：

$$\boldsymbol{PSPA}=\begin{Bmatrix}\boldsymbol{PSPA}_1\\ \vdots\\ \boldsymbol{PSPA}_j\\ \vdots\\ \boldsymbol{PSPA}_N\end{Bmatrix}=\begin{Bmatrix}p_{com1}^{\ 1} & p_{com1}^{\ 2} & \cdots & p_{com1}^{\ 20}\\ \vdots & \vdots & & \vdots\\ p_{com\,j}^{\ 1} & p_{com\,j}^{\ 2} & \cdots & p_{com1}^{\ 20}\\ \vdots & \vdots & & \vdots\\ p_{com\,N}^{\ 1} & p_{com\,N}^{\ 2} & \cdots & p_{com\,N}^{\ 20}\end{Bmatrix} \tag{5}$$

其中，N 表示输入语音段的帧数；$\boldsymbol{PSPA}_j$ 表示第 j 帧基音同步幅度峰值特征矢量；$p_{com\,j}^{\ i}$ 表示第 j 帧经过非线性压缩的第 i 路幅度峰值。

从基音同步幅度峰值特征的提取过程中可以看到，该特征模拟了听觉系统感知语音的特点。PSPA 首先利用 20 个临界带通滤波器组，模拟耳蜗对输入信号的频率选择特性；然后通过全波整流，将信号归整为单边信号；随后根据听觉系统具有基音同步的特点，采用基音同步的方法对信号进行动态分帧；接着根据听觉神经纤维的应激机制，对基音周期内各子带的峰值信号进行非线性压缩；最后，通过 DCT 变换将各频带经过峰值加权频谱域信号变换至倒谱域，得到基音同步幅度峰值特征。从某种意义上说，由于采用了基音同步的方式，因此该特征能够同时表征信号的瞬时频率和强度信息。

在重音检测中，音高和音强(即瞬时频率和强度)均为重要的声学特征，而基音同步幅度峰值特征恰好能够反映这 2 类信息变化的情况，而且利用基音同步的方式巧妙地将这 2 类信息结合在一起考虑。因此，基音同步幅度特征对重音/非重音具有一定区分作用。由于重音信息仅携带在元音音节上，因此上述特征提取方法应根据重音的特点，仅在元音段提取，从而减少运算量。

3 试验与分析

3.1 数据库

本文使用英国利兹大学等录制和标注的 ISLE 英语学习者连续语音数据库。该数据库共收录具有中级英语发音水平的意大利和德国发音者(各 23 名)所朗读的共计 11 484 句英语连续语音数据。语音材料选自 John Hunt 所著《The Ascent of Everest》(5 级英语阅读水平，English Reader Level 5)中的 82 句话共计 1 300 个单词。其中，52 句子中包含 26 个由于重音位置不同导致词性发生变化的动词-名词单词对(如'convict-con'vict)；10 句子中包含由于重音位置改变导致词义出现变化的多音节词(如'photograph-photo'graphic)。ISLE 数据库由 6 位语音专家对所有语音数据在句子、单词和音素 3 个层面进行详细的手工标注，每个句子均由多个语音专家分别进行标注，对存在分歧的标注经讨论给出统一的标注。另外，语音专家还针对重音信息进行了详细的标注，同时，对于英语学习者的发音错误也做了详细的标注。本文选用该数据库中约 1 500 句语音进行试验，在选择语音时做到每个音素出现的比例基本相同。

3.2 系统构成与试验设计

利用基音同步幅度峰值特征进行重音检测的试验框架如图 2 所示。由于 ISLE 数据库已标出自动切分的音段信息，因此利用标注可提取元音段的时长特征；同时提取相应元音段的短时能量特征和基音特征；在提取实时基音信息的基础上根据图 1 所示的方法得到基音同步峰值幅度特征，然后使用 Fisher 线性判据对元音段语音进行重音/非重音检测。

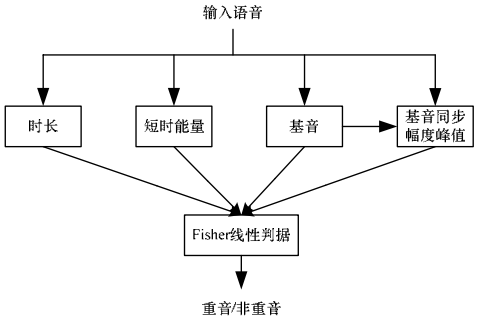

图 2 重音检测流程

为了分析新特征在重音检测中的作用，试验将首先单独使用基音同步峰值幅度、时长、能量和基音等特征分别进行重音检测，通过分析各特征在重音检测性能上的差异，获得各特征对重音检测的贡献度；然后利用新特征与传统特征的组合进行重音检测，分析各特征组合的性能表现，得到重音检测最佳组合。由于重音/非重音的检测是一个二分类问题，因此下文的试验结果均以误识率表示系统的性能优劣。在试验中采取留一法进行交叉验证，以平均误识率作为实验结果。

3.3 试验数据与分析

各类试验结果如表 1 所示，其中，D, E, P 和 PSPA 分别代表时长、能量、基音、基音同步幅度峰值特征。根据表中数据可知：

(1)对于 3 种传统特征，单独使用时长特征获得最低的误识率(21.73%)，单独使用短时能量特征和基音特征的误识率分别为 34.78%和 37.01%。表明 3 类传统特征对重音检测的贡献度依次为时长、能量和基音，这与文献[3]的观点相吻合。

(2)单独使用基音同步幅度峰值特征获得了 22.83%的误识率，这一结果比单独使用时长特征的误识率略高，但比单独使用能量和基音特征都要低。在各特征的组合试验中，联合使用新特征的各类特征组合性能均优于仅使用传统特征的

(下转第 30 页)