基于声学相关特征与词典语法相关特征的汉语重音检测

倪崇嘉[1),2)] 张爱英[1)] 刘文举[2)]

[1)](山东财政学院统计与数理学院 济南 250014)

[2)](中国科学院自动化研究所模式识别国家重点实验室 北京 100190)

摘 要 重音对提高语音合成系统的自然度、可懂度以及语音识别系统的正确率等方面扮演着非常重要的作用.该文基于大规模韵律标注的语料库,利用声学相关特征及词典语法相关特征对汉语重音进行检测.采用 Boosting 集成分类回归树对当前音节的声学相关特征以及词典语法相关特征进行建模,Boosting 集成分类回归树充分利用了当前音节的特性.同时还对词典语法相关特征采用条件随机场方法建模,条件随机场很好地利用了当前音节的上下文特性.最后,将 Boosting 集成分类回归树模型和条件随机场模型加权组合获得识别率更高的混合模型.该混合模型克服了 Boosting 集成分类回归树模型的不足,实现了 Boosting 集成分类回归树和条件随机场的优势互补.实验结果表明该方法具有较好的分类效果,在 ASCCD 语料库上能够获得 84.82%重音检测正确率.同时,与之前其他人的工作在相同的条件下(相同的训练集和测试集)对比,在正确率方面,该方法分别有 4.01%和 1.67%的提高.另外,该文中,对英语的重音检测和汉语的重音检测做了对比,并通过特征分析方法从另一个层面验证了一些语言学上的结论.

关键词 重音;Boosting 集成分类回归树;条件随机场;神经网络;分类回归树

中图法分类号 TP319 **DOI 号**: 10.3724/SP.J.1016.2011.01638

Mandarin Stress Detection Using Acoustic, Lexical and Syntactic Features

NI Chong-Jia[1),2)] ZHANG Ai-Ying[1)] LIU Wen-Ju[2)]

[1)](*School of Statistics and Mathematics, Shandong University of Finance, Jinan* 250014)

[2)](*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract The stress is important to improve the naturalness, understandability and intelligibility of speech synthesis system and the correct rate of automatic speech recognition system. In this paper, we conduct stress detection by using the acoustic, lexical and syntactic features based on large scale prosodic annotation corpus. Boosting classification and regression tree is utilized to model the acoustic, lexical and syntactic features, which adequately utilizes the property of the current syllable. Conditional random fields (CRFs) are utilized to model the lexical and syntactic features, which adequately utilize the contextual property of the current syllable. The combination of boosting classification and regression tree and conditional random fields achieves better classification effect when compared with boosting classification and regression tree model or conditional random fields. The combined model overcomes the efficiency of boosting classification and regression tree model, and realizes the complementarities with the advantages of boosting classification and regression tree and conditional random fields. The experimental results indicate

收稿日期:2009-10-11;最终修改稿收到日期:2011-06-13.本课题得到国家自然科学基金(90820303,60675026,90820011)、国家"八六三"高技术研究发展计划项目基金(20060101Z4073,2006AA01Z194)和国家"九七三"重点基础研究发展规划项目基金(2004CB318105)资助.

倪崇嘉,男,1979 年生,博士,讲师,主要研究方向为语音识别、韵律模型和机器学习.E-mail: cjni@nlpr.ia.ac.cn; nichongjia@gmail.com. **张爱英**,女,1980 年生,硕士,讲师,主要研究方向为机器学习.**刘文举**,男,1960 年生,博士,研究员,博士生导师,主要研究领域为语音识别、语音合成和计算听觉场景分析.

that the proposed method acquires better classification effect, and achieves 84.82% stress detection accuracy rate on ASCCD. Compared with the previous counterpart work in the same conditions (the same training set and testing set), there are 4.01% and 1.67% improvements respectively in terms of the correct rate. In this paper, we also compare the differences and the similarities between Mandarin stress detection and English pitch accent detection. Based on the feature analysis on the large scale prosodic annotation corporus, we also verify some linguistic conclusions in a different way.

Keywords　stress; Boosting classification and regression tree (Boosting CART); conditional random fields (CRFs); neural network; classification and regression tree (CART)

1　引　言

基于大规模语料库，研究语音的韵律特征成为言语工程技术中的研究热点.

韵律是指包含在语音中的节奏(rhythm)、语调(intonation)、强调(accent)等模式，通常通过基频轮廓的变化、音节时长的加长以及强调、重读等表现出来. 韵律在增加语言表达的自然度和可理解度方面扮演着非常重要的作用. 近几十年来的研究成果也表明：韵律特征的引入，能够在减少语音识别的错误率、降低问题的复杂性、增加系统理解的准确性以及提高语音合成的自然度等方面具有非常重要的作用. 基于统计的可训练的韵律模型已经成功地应用于语音合成领域，它对提高语音合成的自然度有很大的帮助. 在语音识别领域，韵律模型已经成功地运用到德语、法语的语音识别[1-4]. 对于汉语，韵律模型也开始逐步地应用到语音识别领域. 但是，由于利用的韵律信息很少，应用的效果并不是很好，特别是在自然口语识别领域. 因此，针对语音识别和语音理解方面的基于大规模语料库的统计韵律模型的研究还需要进一步的深入.

我们知道，汉语具有字、词、词组、短语、句子、段落和篇章等由小到大的单元层级结构. 相应地，汉语的韵律单元也具有类似的层次结构. 尽管目前对汉语韵律层级的定义各有差异，但是大体可以分为如下的层次：音节、音步(韵律词)、韵律短语、语调句子、句群(段落)和篇章. 考虑到言语工程中的实际需要，我们一般只考虑句子层级之内的韵律层级. 关于汉语韵律层级的划分以及韵律划分同句法和语用之间的关系，很多研究者给出了不同的结论. 同时，我们知道在人们进行语言交流时，其交流的语言不仅仅是各个单元的层级结构，而且语言中各个单元的轻重也同样起着非常重要的作用. 重音是说话时为了突出多音节的词、短语或者句子中的某一部分，音发的比其它部分更加有力、突出，因而更加响亮、清楚. 汉语普通话中一般存在两类重音：词重音和句重音. 对于汉语重音，不同的研究人员从不同的角度进行了不同的划分. 鉴于汉语层次单元的边界和汉语重音在声学表现的相关性，我们考虑的是三级重音：韵律词重音、次要韵律短语重音和主要韵律短语以及句子重音. 韵律的层级结构和重音构成了韵律研究中的两个基本的问题.

当前，基于大规模语料库的统计方法对韵律的研究提供了很大的帮助. 为了描述韵律的结构信息、轻重音以及基频运动的模式需要一个统一的框架. 当前，已有很多这样的描述框架，如 ToBI[5]、TILT[6]、Fujisaki 等人[7]、IViE[8]、C-ToBI[9] 等. 但是在构建这些韵律标注的韵律库时，完全依靠人工标注这些韵律结构信息和轻重音是十分耗时的，同时也是不准确的. 并且，也不利于韵律模型在言语工程中的应用. 因此利用计算机，通过建立模型对韵律进行自动标注的研究越来越受到人们的重视.

本文将充分利用来自声学(如基频或音高、能量、音强、时长等方面)的信息以及词典、语法方面的信息，构建汉语重音检测的特征集，采用机器学习的方法对汉语重音进行检测. 同时说明汉语重音的检测与英语重音检测的不同之处，并通过特征分析的方法从另一个层面揭示汉语重音在语言学和语音学上的特点.

本文在第 2 节中将介绍重音检测方面的相关工作；在第 3 节中将对实验用的具有韵律标注的语料库进行介绍；在第 4 节中将列出在汉语重音检测和英语重音检测时所用到的特征，对部分特征进行正则化，并介绍声学-重音模型、语法-重音模型以及它们的加权混合模型；在第 5 节中，对实验环境进行描

述,并对实验结果进行分析;在第 6 节,对实验用到的特征进行分析.最后,将给出我们的结论和下一步的工作.

2 相关工作

针对英语,Wightman 等人[10]提出了韵律模式的自动标注,他们采用决策树和马尔可夫序列模型去判断音节的间断类型和是否重读,具体是在用语音识别系统提供的音节或者音素强制切分信息基础上,抽取音节的时长、基频以及能量的特征,构建决策树,去预测间断的类型和音节是否重读.实验表明:使用计算机对间断类型和音节是否重读进行标注能够较好地与经过专门训练的人标注的结果保持一致,能够达到 83%重音检测正确率.Ananthakrishnan 等人[11]构建了韵律识别系统,采用耦合隐马尔可夫模型(Coupled Hidden Markov Model,CHMM)在音节和词层次上对重音进行检测.CHMM 能够对异质的音强特征、音高特征、时长特征等声学特征和同质的语法特征进行建模,同时又很好地描述了话语的语法结构和韵律结构之间的联系.在波士顿大学广播新闻语料库(Boston University Radio News Corpus,BURNC)上重音检测的实验说明,该识别器在音节层次上能够达到 75%重音检测正确率.而且 Ananthakrishnan 等人[12]2008 年又在上述研究的基础上,在最大后验(Maximum A Prior,MAP)框架下,利用波士顿大学广播新闻语料库对重音检测,能够达到 86.75%的正确率.另外,Ananthakrishnan 等人[13]仅利用 RFC 特征和韵律语言模型对重音进行预测,在 BURNC 语料库上能够达到 67.7%检测正确率.Sridhar 等人[14]在最大熵框架下,利用声学、语法的特征对韵律进行自动标注,其结果是在波士顿大学广播新闻语料库和波士顿 Direction 语料库(Boston Directions Corpus,BDC)上对词的重音分别能够达到 86.0%和 79.8%的正确率.Johnson 等人[15]利用神经网络和高斯混合模型在 BURNC 语料库上对词的重音能够达到 84.2%检测正确率.Rosenberg 等人[16]实验了在2-20bark 上,仅采用能量相关特征,利用分类回归树(Classification And Regression Tree,CART)的 C4.5 算法对重音进行检测.通过对 2-20bark 频带上构建的总共 210 个 CART 分类器进行投票集成,在 BDC 语料库上对重音的检测正确率能够达到 81.9%.并且其在 Interspeech 2007 上发表的文章又结合音高特征,同时还是采用上述类似的方法构造 210 个不同的 CART 分类器进行投票集成,在朗读语音、自然语音和广播新闻语音语料库上,对重音的检测正确率分别能够达到 84.0%、88.3%和 88.5%[17].Sun[18]利用 Pitch Target 特征,同时结合时长、能量以及一些文本特征构造特征集,采用集成机器学习的 Boosting 和 Bagging 方法训练分类器,分别能够达到 87.17%和 84.26%的重音检测正确率.Hun 等人[19]利用声学和词典语法方面的信息,采用神经网络和支持向量机方法建立声学—重音模型和语法—重音模型,并通过加权的方法对声学—重音模型和语法—重音模型进行集成,实验表明该混合模型在音节层次上对重音能够达到 89.84%的检测正确率.Margolis 等人[20]利用 Boosting 方法、决策树以及高斯线性分类器对重音进行检测,分别达到 88.0%±1.8%,86.3%±1.7%,87.1%±1.4%的检测正确率.

针对汉语的重音检测,胡伟湘等人[21]利用时长、音高、能量以及文本相关特征,利用自动感知概率混合模型,在韵律标注语料库 ASCCD 上分别能够达到 84.2%和 81.0%的重音检测正确率.邵艳秋等人[22]也对汉语重音检测进行了研究,他们利用神经网络分别对声学相关特征以及语法相关特征建立声学模型、语法模型以及混合模型,并利用混合模型对轻声、正常重音和重读进行判别,达到了较单一的声学-重音模型或语法-重音模型更好的 84.3%检测正确率.

综合上述重音检测方法,可以将其分为两类:(1)分别对声学相关特征和语法相关特征建模,然后通过加权组合来获得更好的分类器;(2)直接对所有的特征进行建模.第 1 类方法不足之处在于:虽然最后通过加权的方法刻画了声学相关特征以及词典语法相关特征之间的关系,但是它们之间的更深层次联系没有被分类器所利用.从模型的层次上来讲,这类方法仅仅利用了当前音节所提供的特征.第 2 类方法不足之处在于:虽然利用了来自声学和词典语法相关特征训练模型,强化了声学和语法相关特征之间的联系,但是没有重点突出声学或者词典语法方面的特征,更没有很好地在模型层次上利用上下文特征.我们的方法在很大程度上克服了这些不足,充分利用了来自声学以及词典语法等方面的信息,采用 Boosting CART 方法对所有特征进行建模,同时又对词典语法相关的特征采用条件随机场(CRFs)进行建模,最后,通过加权的方法对这两种

模型进行组合. 由于 Boosting CART 方法不仅很好地反映了当前音节属性而且又在更深层次上反映属性之间的联系，同时 CRFs 又能够很好地反映该音节的上下文特性，这种在模型层次上的互补特性使得加权以后的模型获得了较好的识别效果. 与之前的方法在同样条件下的对比，即在相同的训练集和测试集上进行对比表明，我们提出的方法分别有 4.01%和 1.67%的提高.

3　韵律标注语料库

具有韵律标注的语料库 ASCCD 是标准普通话流畅话语语料库. 该语料库由中国社会科学院语言研究所语音实验室收集录音并准确标注. 语料文本是 18 篇叙事体、议论体语篇，每篇 3～5 个自然段，每个自然段 500～600 个音节，总计约 9000 个音节，共 10 个发音人，5 男 5 女，分别记为 M001、M002、M003、M004、M005、F001、F002、F003、F004、F005，用标准普通话，以自然的方式，适中语速，流畅地朗读语篇. 所有语音都经过标注，音段采用 SAMPA-C 标准标注[23]，韵律采用 C-ToBI 韵律标注系统标注，其标注了音节拼音、声韵母、声调、韵律边界等级以及语句重音信息[9]. 其中重音用 1、2、3 分别表示韵律词重音、次要韵律短语(MIP)重音和主要韵律短语(MAP)重音，0 则表示不重读，即正常读音. 在本文中，我们将音节分为轻声、正常读音和重音，而不去区分重音之间的差别，将韵律词重音、次要韵律短语(MIP)重音和主要韵律短语(MAP)重音都看作是重读. ASCCD 语料库中重音的分布如表 1 所示.

表 1　ASCCD 语料库中重音的分布

读法	音节数目	所占百分比
全部	87586	100.00
轻声	5509	6.29
正常读音	48147	54.97
重音	33930	38.74

波士顿大学广播新闻语料库(Boston University Radio News Corpus，BURNC)是具有韵律标注的广播新闻朗读风格的英语语音库[24]. BURNC 对来自 3 个女声(f1a，f2b 和 f3a)和 3 个男声(m1b，m2b 和 m3b)大约 3 个小时的语料采用 ToBI 标准进行韵律标注. 主要标注了重音、语调短语边界(Intonational Phrase Boundary，IPB)、间断(break)等韵律信息. 同时，该语料库利用识别器对相应的语音文本进行了强制切分，提供了音子、音节、单词的时间边界信息. 另外，该语料库还提供了词的词性、音高等标注信息. BURNC 语料库中重音分布如表 2 所示.

表 2　BURNC 语料库中重音的分布

	女声重音分布			男声重音分布		
	f1a	f2b	f3a	m1b	m2b	m3b
语句	74	164	33	72	51	24
词	3993	12607	2733	5059	3608	2093
重音	2344	7061	1545	2786	2113	1094

4　基于声学相关及词典、语法相关特征的加权组合的重音识别

4.1　汉语重音检测时声学及词典、语法相关特征

4.1.1　声学相关特征

文献[25-26]表明，声学特征，如时长、能量和基频，与重音有很强的相关性，并且一些词典语法相关特征，如词的词性等也与重音有很强的联系. 因此，在本文中我们将采用时长、基频、能量和音强相关特征对汉语重音进行预测. 同时，为了消除不同说话人的影响，我们对部分特征采用 Z-SCORE 算法进行正则化.

对于时长特征，为每一个音节计算下列特征：

(1) DurS. 当前音节的时长；

(2) NSilD. 当前音节后无声段的时长；

(3) PSilD. 当前音节前无声段的时长；

(4) NSilType. 当前音节后无声段的类型；

(5) PSilType. 当前音节前无声段的类型；

(6) bDurType. 当前音节的时长在韵律词范围内是否是最长的；

(7) NormalDur. 当前音节的被 Z-SCORE 算法正则化后的时长.

同时我们还计算当前音节与之前音节时长以及之后时长的比值，并计算与音节的韵母相关的时长特征.

对于音高，为每一个音节计算如下统计特征：

(1) PMax. 当前音节音高的最大值；

(2) PMin. 当前音节音高的最小值；

(3) PMean. 当前音节音高的平均值；

(4) PRMS. 当前音节音高的均方根；

(5) PStddev. 当前音节音高的标准差；

(6) bPthType. 当前音节的音高的最大值是否是韵律词范围内的最大值的表征.

类似于基频相关特征，我们可以计算能量、音强相关的统计特征. 同时，考虑到音节上下文的影响，

我们在当前音节的上下文窗口中计算音高、能量和音强相关的动态特征，并正则化. 设 $E^C_{\max}$、E^C_{mean} 分别表示当前音节范围内的基频或能量的最大值和平均值，$E_{\max}$、$E_{\min}$、E_{mean}、$E_{\text{std.dev}}$ 分别表示在上下文窗口中基频或能量的最大值、最小值、平均值和标准差. 则按照如下的方法计算音高、能量和音强的动态特征.

(1) 分别计算在 3 个音节范围内的均值，对均值进行直线拟和，得到斜率；

(2) $\dfrac{E^C_{\max}-E_{\text{mean}}}{E_{\text{std.dev}}}$；

(3) $\dfrac{E^C_{\text{mean}}-E_{\text{mean}}}{E_{\text{std.dev}}}$；

(4) $\dfrac{E^C_{\max}-E_{\max}}{E_{\text{std.dev}}}$；

(5) $\dfrac{E^C_{\max}}{E_{\max}-E_{\min}}$；

(6) $\dfrac{E^C_{\text{mean}}}{E_{\max}-E_{\min}}$.

考虑到汉语中单音节词和双音节词所占的比重较高，并且在重读时，当前音节之前的音节对重音的影响程度要大于当前音节之后的音节，因此，选择当前音节之前的两个音节以及之后的一个音节作为当前音节的上下文窗口. 并且之前的关于英语重音和荷兰语重音的研究表明：在 500Hz 到 2000Hz 频带上的能量与重音有着密切的联系[16,27]. 因此，在计算能量相关特征时，只是计算在 500Hz 到 2000Hz 频带上的能量.

4.1.2 词典、语法相关特征

对于文本特征，首先我们利用 Stanford 中文分词工具[28-29] 和 Stanford 中文词性标注工具[30] 对文本进行分词和词性标注，获得分词信息和词性标注信息，然后对每一个汉字，计算下列词典语法相关特征：

(1) BSeg. 当前音节是否是分词的边界；

(2) PSum1. 当前音节距之前的韵律边界的音节个数；

(3) NSum1. 当前音节距之后的韵律边界的音节个数；

(4) SSum1. 当前音节距句首的音节距离；

(5) SSum2. 当前音节距句尾的音节距离；

(6) Tag. 当前音节的词性；

(7) PTag. 当前音节前音节的词性；

(8) NTag. 当前音节后音节的词性；

(9) AccentRatio. 当前音节在训练语料库中重读的频率(概率)；

(10) Joint. 在当前音节之前的音节与当前音节被重读的频率(概率)；

(11) ReJoint. 在当前音节之后的音节与当前音节被重读的频率(概率)；

(12) Bigram. 在当前音节之前的音节给定的情况下，当前音节被重读的频率(概率)；

(13) RevBigram. 在当前音节之后的音节给定的情况下，当前音节被重读的频率(概率)；

(14) BBk. 当前音节的韵律间断层级；

(15) T. 当前音节的音调；

(16) T1. 当前音节之前音节的音调；

(17) T2. 当前音节之后音节的音调；

(18) Wlen. 当前音节所在的词的长度；

(19) WSum1. 当前音节距所在词开始的音节距离；

(20) WSum2. 当前音节距所在词结束的音节距离；

(21) PosC. 当前音节在韵律词中位置，分为词首、词中和词尾 3 种；

(22) WPosC. 当前音节在词中的位置，分为词首、词中和词尾 3 种；

(23) PPosC. 当前音节所在的韵律词在韵律短语中的位置，分为词首、词中和词尾 3 种.

4.2 英语重音检测时声学及词典、语法相关特征

为了验证我们的方法，在英语韵律标注语料库 BURNC 上，我们也进行了实验. 对于每一个词，我们计算如下来自声学、词典以及语法方面的特征.

对于音高，计算下列特征：

(1) PMax. 音高的最大值；

(2) PMin. 音高的最小值；

(3) PMean. 音高的平均值；

(4) PRange. 音高的范围；

同时计算音高曲线的轮廓特征：

(5) Pth_a_i，$i=0\sim5$. 音高曲线的 5 阶 Legendre 多项式展开系数.

对于能量，计算下列特征：

(1) EngMax. 能量的最大值；

(2) EngMin. 能量的最小值；

(3) EngMean. 能量的平均值；

(4) EngRange. 能量的范围；

同时计算能量曲线的轮廓特征：

(5) Eng_ a_i，$i=0\sim5$. 能量曲线的 5 阶 Legendre 多项式展开系数.

对于时长，为每一个词，计算时长：

DurWrd. 该词的时长.

对于每一词，计算词典和语法相关特征：

(1) WordID. 词的 ID；

(2) bLexicalStress. 该词是否存在词典重音；

(3) POSTag. 词的词性标注信息.

对于能量或音高曲线的 5 阶 Legendre 多项式展开系数的计算方法如下：

假设 $f(t)$表示能量或音高曲线，t 表示时间，则 $f(t)$的 Legendre 多项式展开为

$$f(t)\approx\sum_{n=0}^{M}a_nP_n(t) \tag{1}$$

其中，

$$P_n(t)=\begin{cases}1, & n=0\\ t, & n=1\\ \dfrac{2n-1}{n}tP_{n-1}(t)-\dfrac{n-1}{n}P_{n-2}(t), & n\geq 2\end{cases}$$

是第 n 阶 Legendre 多项式. a_n 是 $f(t)$第 n 阶 Legendre多项式展开式的系数.

同时，我们在由当前词之前的两个词以及之后的两个词所组成的上下文窗口中，计算词典和语法特征.

4.3 基于声学-重音模型和语法-重音模型加权组合重音识别

设 $W=\{w_1,w_2,\cdots,w_n\}$ 是音节序列，$A=\{a_1,a_2,\cdots,a_n\}$是相应的声学-重音特征序列，$S=\{s_1,s_2,\cdots,s_n\}$是相应的语法-重音特征序列，那么 W 的最有可能的重音标注序列 P^* 可以表示为

$$P^*=\arg\max p(P|A,S) \tag{2}$$

$$\approx\arg\max p(P|A)p(P|S) \tag{3}$$

$$\approx\arg\max\prod_{i=1}^{n}p(p_i|a_i)^{\lambda}p(p_i|\phi(s_i)) \tag{4}$$

$$\approx\arg\max\lambda\sum_{i=1}^{n}\log(p(p_i|a_i))+\sum_{i=1}^{n}\log(p(p_i|\phi(s_i))) \tag{5}$$

其中，$\log(p(p_i|a_i))$是声学-重音模型的得分，$\log(p(p_i|\phi(s_i)))$是语法-重音模型的得分，λ 是区分不同方法建立模型的权重. 我们采用了不同的方法建立声学-重音模型和语法-重音模型，例如分类回归树(Classification And Regression Tree，CART)、神经网络(Neural Network，NN)、支持向量机(Support Vector Machine，SVM)、条件随机场(Conditional Random Fields，CRFs)、决策树(Decision Tree，DT)等方法进行建模.

对于语法-重音模型，我们采用条件随机场对词典语法相关特征进行建模. 条件随机场(Conditional Random Fields，CRFs)是一个无向图模型，CRFs 模型经常用于序列数据的标注，被广泛应用到自然语言处理中[31].

在利用 CRFs 建模时，要对连续的数据进行离散，我们采用将区间等分成 10 份的 bin 方法. 虽然可以等分成更多的份数，但是我们发现其分类效果和分成 10 份的类似，因此，我们的实验只是分成 10 份的实验结果.

对于利用全部特征训练得到的 Boosting CART 模型和仅利用语法特征训练得到的 CRFs 模型，我们利用式(5)进行加权融合.

5 实验以及实验结果分析

5.1 实验环境

在汉语语料库 ASCCD 上，我们从每个说话人的 75 句话中，随机选择 50 句作为训练集，其它的 25 句作为测试集，在句子层次上训练集与测试集的大小是 2∶1，在音节层次上训练集共包含了 58 949 个音节，测试集上共包含了 28 637 个音节. 对于分类回归树模型，我们采用 WEKA 的 C4.5 算法和 WEKA 的默认设置训练得到. 对于神经网络模型，我们采用 WEKA 的多层感知器(Multi-Layer Perception，MLP)，设置了 1 个隐层，隐层所含节点的个数等于输入特征的一半. 对于支持向量机模型，我们采用 WEKA 的 SMO 分类器以及其默认设置[32]. 我们利用 WEKA 中的 MultiBoostAB 作为强分类器，分类回归树(Classification And Regression Tree，CART)作为弱分类器，训练 Boosting CART 模型. 对于 CRFs 模型，我们采用 CRF++0.53 工具训练得到[33].

在英语语料库 BURNC 上，我们采用 5 折交叉验证(5-fold cross validation)来验证我们的方法，其最后的实验结果是 5 折交叉验证结果的平均. 模型训练的工具选择和设置与 ASCCD 语料库上训练相应模型时的选择和设置是一样的.

5.2 实验结果及分析

表 3 列出了分别利用分类回归树和神经网络对声学特征进行建模获得的声学-重音模型的重音检测结果. 从表 3 可以看到，分类回归树模型和神经网络模型的分类效果差不多，但是基于神经网络的声学-重音模型的分类效果稍好.

表 3 不同的声学-重音模型的重音检测结果

方法	平均正确率/%
分类回归树(CART)	77.26
神经网络(Neural Network,NN)	77.48

表 4 列出了基于分类回归树(CART)、支持向量机(SVM)以及条件随机场(CRFs)的语法-重音模型的重音检测结果.从表 4 可以看到,基于 CRFs 的语法-重音模型由于很好地刻画了上下文对当前音节的影响,从而获得了最好的结果.支持向量机的识别结果最差.

表 4 不同的语法-重音模型的识别性能

方法	平均正确率/%
分类回归树(CART)	81.27
支持向量机(SVM)	78.81
条件随机场(CRFs)	81.54

表 5 列出了不同的声学-重音模型和语法-重音模型通过式(5)加权组合得到的模型的重音检测的正确率.在表 5 中,Boosting CART 模型不是采用式(5)加权获得,而是直接利用所有的特征训练获得.从表 5 可以看出:(1)通过将声学-重音模型和语法-重音模型加权获得的混合模型的识别性能没有通过集成方法获得的识别器的性能好;(2)利用 CRFs 对词典语法相关特征建模获得的模型比利用 SVM 对词典语法相关特征建模获得模型有更好的识别性能;(3)利用 NN 对声学相关特征得到的识别器的性能要比用 CART 对声学相关特征得到的识别器的性能要差,而在表 3 中,基于 NN 的声学-重音模型的识别率要比基于 CART 的声学-重音模型的识别率要高.其原因是:基于 NN 的声学-重音模型与语法-重音模型识别器的重叠部分要比基于 CART 的声学-重音模型与语法-重音模型的重叠部分要多,从而使得当基于 NN 的声学-重音模型与语法-重音模型结合时所得到的分类器的性能要稍逊于基于 CART 的声学-重音模型与相应的语法-重音模型结合时所得到的分类器.

表 5 不同声学-重音模型和语法-重音模型的混合模型的识别性能

混合模型	平均正确率/%
NN/SVM	80.50
NN/CRFs	82.78
CART/SVM	80.69
CART/CRFs	83.14
Boosting CART	84.59

最后,我们利用加权的方法对利用全部特征获得的 Boosting CART 模型与仅利用词典语法特征获得的 CRFs 模型进行组合.同样地,我们也对 NN/SVM、NN/CRFs、CART/SVM、CART/CRFs 混合模型分别与基于 CRFs 和 SVM 的语法-重音进行加权再组合,其识别的结果在表 6 列出.从表 6 可以看出,NN/SVM、NN/CRFs、CART/SVM、CART/CRFs 与基于 CRFs 或 SVM 的语法-重音模型进行加权再组合时,识别的平均正确率同没有与基于 CRFs 或 SVM 的语法-重音模型进行结合之前的混合模型的结果相比,识别的正确率上下稍微有点波动.其原因应该有两个:(1)当之前的混合模型再与同样的语法-重音模型结合时,由于我们实验的时候权重是按照 0.1 的间隔进行增加或减少的,当它们再进行结合时,有的权重在之前没有出现过,所以会有识别率增加或减少现象发生;(2)当之前的混合模型再与不一样的语法-重音模型结合时,如果之前的声学-重音模型和一个与现在的语法-重音模型相比较差的一个语法-重音模型结合,现在再与语法-重音模型进行结合时,识别的正确率会有所提高的.这是由于之前的有些信息没有被充分的挖掘出来,现在遇到一个比之前结合的语法-重音模型要好的语法-重音模型时,原来没有被发现的信息又被挖掘,从而使得识别率提高.当然,如果之前声学-重音模型结合的语法-重音模型较现在的语法-重音模型要好,那么在模型的限制下,其已经达到了它们的最优状态,信息已经被充分的挖掘,那么再进行结合一个较弱的语法-重音模型时,当前的识别就会被看作噪声的干扰,会使得识别率下降.我们在实验过程中的权重的走势也说明了这一点.

表 6 加权混合模型与基于 CRFs、SVM 的语法-重音模型进行加权再组合后的结果

组合模型	平均正确率/%
NN/SVM+SVM	80.54
NN/SVM+CRFs	82.57
NN/CRFs+SVM	82.45
NN/CRFs+CRFs	82.35
CART/SVM+SVM	80.88
CART/SVM+CRFs	82.90
CART/CRFs+SVM	83.30
CART/CRFs+CRFs	83.14
Boosting CART+SVM	84.10
Boosting CART+CRFs	84.82

从表 6 我们看到,当利用来自声学以及词典语法相关的所有特征建模获得的 Boosting CART 模型再与基于 SVM 和 CRFs 的语法-重音模型再进行结合时,其识别性能一个是下降,另外一个是上升.这是十分有趣的现象.那么什么原因导致了这种情况出现呢?最根本的原因是分类器的类型.我们知

道，SVM 分类器只是考虑了当前的音节的特征，也就是我们所提供的声学和语法相关的特征，没有考虑其上下文的特征. 而 CRFs 就不同了，虽然我们只是采用了一阶的情况，但是 CRFs 在模型层次上利用了当前音节的上下文特征. 虽然，我们在构造声学和词典语法相关特征时，也考虑当前音节在上下文窗口中的特征，但是，此时的 CRFs 所利用的上下文特征和它们是不同的. 我们知道，由于仅利用当前音节的所有输入的特征进行模型训练时，Boosting CART 方法已经很好地利用了这些特征，构造了一个较好的分类器，因此当一个较之差的分类器再与它进行结合时，分类性能势必会下降. 当 Boosting CART 与 CRFs 进行结合时，由于模型的互补特性，它们的结合不仅考虑了当前音节的属性，同时也考虑了当前音节的上下文属性，从而使得其分类的性能提高.

为了更好地验证我们所提的方法，我们与之前的方法进行了比较. 我们引入文献[21]的在 ASCCD 语料库上的训练集 Tr1 和 Tr2 以及测试集 T1 和 T2. Tr1 是由说话人 M001 的前 12 句话组成，T1 是由说话人的最后 6 句话组成. Tr2 是由所有的说话人的前 12 句话组成，T2 是由所有说话人的最后 6 句话组成的. 实验的结果在表 7 中列出. 从表 7 可以看出，我们的方法较文献[21]的结果在测试集 T1 和 T2 上正确率分别提高了 4.01% 和 1.67%. 我们的模型很好地结合了当前音节及其上下文的属性，较好地融合了声学相关信息和词典语法相关信息，获得了不错的检测效果.

表 7　不同方法之间的检测结果的比较

	T1 上的正确率/%	T2 上的正确率/%
我们的方法	88.21	82.67
文献[21]的结果	84.20	81.00

5.3　方法的进一步验证

为了更进一步验证我们所提的方法，我们又在英语的 BURNC 语料库上进行了实验. 表 8 列出了不同的模型在 BURNC 语料库上的分类效果.

表 8　不同的模型在 BURNC 语料库上的重音检测结果

模型	平均正确率/%
NN/CART	84.4
NN/SVM	83.7
NN/CRFs	86.5
Boosting CART	85.2
CRFs	86.7
Boosting CART+CRFs	87.7

从表 8 我们可以看到：我们所提的方法能够获得最高的 87.7% 的重音检测正确率. 同时，对比表 6 和表 8，我们的方法在 ASCCD 语料库上能够获得 84.82% 的检测正确率，而在 BURNC 语料库上能够获得 87.7% 的检测正确率. 从数值上讲，ASCCD 语料库上获得的正确率比 BURNC 语料库上获得的正确率要低一些. 将我们的实验结果与文献[10,12,14]的结果进行对比，发现我们的方法获得了较好的分类效果.

5.4　实验结果的进一步分析

文献[19]对英语重音(pitch accent)检测的不同结果、方法进行了对比分析，得到结论：对于声学-重音模型，采用 3 层的神经网络且设置隐层所含节点个数等于输入特征一半时获得较好的分类效果；而对于语法-重音模型，采用 SVM 进行建模且建模时采用 WEKA 的默认设置时达到较好的分类效果. 我们采用该方法对汉语的重音进行检测，其结果在表 3～表 7 列出. 从这 5 个表中我们可以看到，汉语重音的检测的正确率要低于一般英语重音的检测的正确率.

汉语是单音节的声调语言. 声调有非常重要的辩义作用. 在汉语中，声调主要是通过音高曲线的变化来实现. 而汉语的音高曲线负载了声调、重音和语调等信息. 因此，音高曲线不能自由地标识重音. 英语是重音的语言. 英语中该词是否具有词典重音与该词是否重读有很大的关系.

对于 ASCCD 语料库，文献[21]对重音的分布进行了统计，从其统计的表中可以看到：(1) 轻声是比较稳定的，也就是说 10 个说话人在轻声的模式上是基本相同的. 在语料库中，被所有说话人都读做轻声的音节共有 722 个音节，被所有说话人正常读而没有重读的音节有 1174 个，被所有说话人都重读的音节仅为 682 个. 而表 1 表明，整个语料库中共有 33930 个重读的音节，平均到每个人有 3393 个重读的音节. 也就是说对于每一个说话人大约有 2711 个是可以自由地重读或正常读. 而每一个人总共读了大约 8758 左右的汉字. 对每一个人来说，可以重读或正常读的这一部分音节占其所读汉字总数的 30% 左右. 可见，汉语中对汉字是否重读的自由度很大，这也决定了汉语重音检测是一件很困难的事情. 在英语 BURNC 语料库中，重音标注的一致性是很高的. 首先，BURNC 是采用 ToBI 框架进行标注的. ToBI 标注系统的一个优点就是：对于不同的风格，

它都能够一致地表示[34]. 其次,Ostendorf 等人对 BURNC 语料库标注的一致性进行了统计研究. 在 BURNC 语料库中包含 1002 个词由 3 个小故事组成的一个集合上,她们研究发现有 487 个词被两个不同的韵律标注组标注为存在重读,标注的一致性达到 60%. 而标注最不一致的地方是"L+H*"和"H*"之间的混淆. 如果按照文献[34]中那样,将"L+H*"和"H*"分为一组,重音标注的一致性能够达到 81%[24].

6 特征分析

Sluijter 等人[27]认为对英语来说,时长相关特征与重音有着非常可靠的联系,其次是频谱斜度(spectral balance). 音强相关特征以及元音品质提供了最弱的区分性. Fry 等人[35]通过实验也说明了时长相关特征对重音感知的作用远远超过音强相关特征. 而后来 Bolinger[36]的实验表明,音高的凸现要比时长的改变对重音的影响大. 汉语与英语不同,汉语是一种带调的语言. 声调带有非常重要的辨义作用. 这样,汉语的基频曲线不仅负载了声调,而且还负载了重音、语调等混合信息,汉语的基频曲线不能自由地表示重音. 这使得汉语重音检测非常困难. 那么对汉语来说,哪种才是影响汉语重音感知最重要的因素? 不同的人给出了不同的答案. 赵元任先生认为,汉语重音特征表现为音域加宽、音程加大,其次才是气流加强[37]. 林茂灿等人[38]认为,汉语重音最重要的特征是音长增加,而音强的作用不是想象中那么大. 沈炯等人[39]则认为,在听辨重音时,时长的作用并不明显,而音高的作用很重要. 在本部分中,我们将通过大词汇量韵律标注的语料库对时长、音高、基频以及音强在重音感知中的作用进行分析,从另一个侧面来验证已有的语言学和语音学上的关于重音感知的结论.

6.1 不同的特征组

我们从综合的角度衡量时长、基频、能量、音强以及词典语法相关特征对重音感知的作用. 首先,我们分别利用时长、基频、能量、音强以及词典语法特征在训练集上训练模型,然后利用这些训练好的模型在测试集上去测试这些模型的性能. 表 9 列出了 ASCCD 语料库上,不同的特征组在重音检测中的分类效果.

表 9 ASCCD 语料库上不同的特征组在重音检测中的分类效果

特征组	平均正确率/%
时长	74.6
基频	73.5
能量	67.4
音强	67.3
词典和语法	82.8

从表 9 可以看到:(1)对汉语重音检测来说,词典语法特征在汉语重音检测中起到特别重要的作用. (2)对汉语的重音检测来说,声学相关特征对重音检测的贡献不是很大. 这也说明了,汉语重音感知是十分复杂的现象,是来自各方面特征的综合实现. (3)对汉语的重音检测来说,在来自时长、基频、能量以及音强方面的声学特征中,时长相关特征对重音有很好的区分性,其次是基频、能量、音强相关特征. 这也从另一个侧面说明了在汉语重音感知中,时长的作用是最明显的. (4)在 500~2000Hz 之间的能量虽然对英语和荷兰语的重音有很好的区分性,但是在汉语重音检测中所起的作用不是很明显.

6.2 单个特征

我们从假设检验的角度来衡量单个特征对重音感知的作用. 在训练集上,我们利用方差不同两样本 T 检验方法计算不同特征均值之间的差值,把计算得到的置信度由高到低排序.

对于时长相关特征,bDurType(当前音节的时长在韵律词范围内是否是最长的)、DurS(当前音节的时长)以及当前音节与之前音节时长之比排在前三位. 音节没有正则化的时长 DurS 对重音的区分性要比音节正则化的时长 NormalDur 对重音的区分性要好,这一个现象在英语中也有[19]. 对于基频相关特征,bPthType(当前音节的音高的最大值是否是韵律词范围内的最大值)、PMax(当前音节音高的最大值)以及 PMean(当前音节音高的平均值)这 3 个特征排在前 3 位. 对于能量相关特征,能量的最大值、能量的均值以及能量的最大值在上下文窗口中动态变化,这 3 个特征排在前 3 位. 对于音强特征,音强在音节内的标准差、音强均值是否是韵律词内最大值以及音强在音节内的最大值,这 3 个特征排在前 3 位.

对于词典和语法相关特征,我们发现,与韵律词相关的特征、概率相关的特征以及音节在词中的位置相关的特征,这三类特征对汉语重音的检测特别重要. 概率相关的特征在重音检测中很重要,这一个特点英语中也有类似的发现[26]. 与韵律词相关的特

征在汉语重音检测中很重要，我们认为是与我们采用的重音定义方式有直接的联系. 在 ASCCD 语料库所采用的重音标注框架下，在韵律词中，规定有且仅有一个音节被重读. 而音节在词中的位置相关的特征在重音检测中很重要，这一特点很早就被语言学家所认识[38].

7　总结及展望

本文基于大规模韵律标注的语料库，利用声学相关特征以及词典语法相关特征对汉语重音进行检测. 采用 Boosting CART 对当前音节的声学以及词典语法相关特征建模，该方法充分利用了当前音节的相关特性. 同时我们还对词典语法相关特征采用条件随机场方法建模，条件随机场很好地利用了当前音节的上下文特性. 最后将 Boosting CART 模型和条件随机场模型加权组合获得识别率更高的混合模型. 实验的结果表明该方法具有良好的分类效果. 同时，我们对英语的重音检测和汉语的重音检测做了对比，并通过特征分析方法从另一个层面揭示了汉语重音在语言学和语音学上的特点. 将来，我们要对所用的特征进行简化，并探索其它的建模方法和技术以刻画重音的属性.

参　考　文　献

[1] Gallwitz F, Batliner A, Buckow J et al. Integrated recognition of words and phrase boundaries//Proceedings of the International Conference on Spoken Language Processing. Sydney, Australia, 1998: 2883-2886

[2] Hirschberg Julia, Swerts Marc. Prosodic cues to recognition errors//Proceedings of the Automatic Speech Recognition and Understanding Workshop. Keystone, 1999: 345-352

[3] Hirschberg Julia, Litman Diane, Swerts Marc. Generalizing prosodic prediction of speech recognition errors//Proceedings of the International Conference on Spoken Language Processing. Beijing, China, 2000: 615-618

[4] Hirschberg Julia. Communication and prosody: Functional aspects of prosody. Speech Communication, 2002, 36(1-2): 31-43

[5] Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J. ToBI: A standard for labeling English prosody//Proceedings of the International Conference on Spoken Language Processing. Banff, Alberta, Canada, 1992: 867-870

[6] Taylor P. The TILT intonation model//Proceedings of the International Conference on Spoken Language Processing Sydney. Australia, 1998, 4: 1383-1386

[7] Fujisaki H, Hirose K. Modeling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation//Proceedings of the International Congress of Linguistic. Tokyo, Japan, 1982: 57-70

[8] Grabe E, Nolan F, Farrar K. IViE—A comparative transcription system for international variation in English//Proceedings of the International Conference on Spoken Language Processing. Sydney, Australia, 1998: 1259-1262

[9] Li Aijun. Chinese prosody and prosodic labeling of spontaneous speech//Proceedings of the Speech Prosody 2002. Aix-en-Provence, France, 2002: 39-46

[10] Wightman C W, Ostendorf M. Automatic labeling of prosodic patterns. IEEE Transactions on Speech Audio Process, 1994, 2(4): 469-481

[11] Ananthakrishnan S, Narayanan S. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, PA, USA, 2005: I-269-I-272

[12] Ananthakrishnan S, Narayanan S. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Transactions on Audio, Speech, and Language Process, 2008, 16(1): 216-228

[13] Ananthakrishnan S, Narayanan S. Fine-grained pitch accent and boundary tone labeling with parametric F0 features//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, Nevada, USA, 2008: 4545-4548

[14] Sridhar V K R et al. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Transactions on Audio, Speech, and Language Process, 2008, 16(4): 797-811

[15] Johnson M H et al. Simultaneous recognition of words and prosody in Boston University radio speech corpus. Speech Communications, 2005, 46(3-4): 418-438

[16] Rosenberg A, Hirschberg J. On the correlation between energy and pitch accent in read English speech//Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006-ICSLP). Pittsburgh, Pennsylvania, USA, 2006: 201-204

[17] Rosenberg A, Hirschberg J. Detecting pitch accent using pitch-corrected energy-based predictors//Proceedings of the Interspeech. Antwerp, Belgium, 2007: 2777-2780

[18] Sun Xuejing. Pitch accent prediction using ensemble machine learning//Proceedings of the International Conference on Spoken Language Processing. Denver, Colorado, USA, 2002: 953-956

[19] Hun J, Liu Y. Automatic prosodic events detection using syllable-based acoustic and syntactic features//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Taipei, Taiwan, China, 2009: 4565-4568

[20] Margolis A, Ostendorf M. Acoustic-based pitch-accent detection in speech: Dependence on word identity and insensitivity to variations in word usage//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Taipei, Taiwan, China, 2009: 4513-4516

[21] Hu Wei-Xiang, Dong Hong-Hui, Tao Jian-Hua, Huang Tai-Yi. Study on stress perception in Chinese speech. Journal of Chinese Information Processing, 2005, 19(6): 78-83(in Chinese)
(胡伟湘, 董宏辉, 陶建华, 黄泰翼. 汉语朗读话语重音自动分类研究. 中文信息学报, 2005, 19(6): 78-83)

[22] Shao Yan-Qiu, Han Ji-Qing, Liu Ting, Zhao Yong-Zhen. Study on automatic prediction of sentential stress with natural style in Chinese. Acta Acustica, 2006, 31(3): 203-210 (in Chinese)
(邵艳秋, 韩纪庆, 刘挺, 赵永贞. 自然风格言语的汉语句重音自动判别研究. 声学学报, 2006, 31(3): 203-210)

[23] Chen Xiao-Xia, Li Ai-Jun, Sun Guo-Hua, Wu Hua, Yin Zhi-Gang. An application of SAMPA-C for standard Chinese//Proceedings of the International Conference on Spoken Language Processing. Beijing, China, 2000: 652-655

[24] Ostendorf M, Price P J, Shattuck-Hufnagel S. The Boston University Radio News Corpus: Linguistic Data Consortium, 1995

[25] Pitrelli J F. ToBI prosodic analysis of a professional speaker of American English//Proceedings of the Speech Prosody. Nara, Japan, 2004: 557-560

[26] Nenkova A, Brenier J, Kothari A et al. To memorize or to predict: Prominence labeling in conversational speech//Proceedings of the HLT-NAACL. Rochester, NY, USA, 2007: 9-16

[27] Sluijter A M C, van Heuven V J. Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America, 1996, 100(4): 2471-2485

[28] Tseng H, Chang P, Andrew G, Jurafsky D, Manning C D. A conditional random field word segmenter//Proceedings of the 4th SICHAN Workshop on Chinese Language Processing. Jeju Island, Korea, 2005: 168-171

[29] Chang P C, Galley M, Manning C D. Optimizing Chinese word segmentation for machine translation performance//Proceedings of the ACL 3rd Workshop on Statistical Machine Translation. Prague, Czech Republic, 2008: 224-232

[30] Toutanova Kristina, Klein Dan, Manning Christopher, Singer Yoram. Feature-rich part-of-speech tagging with a cyclic dependency network//Proceedings of the HLT-NAACL. Edmonton, Canada, 2003: 173-180

[31] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the International Conference on Machine Learning. Williamstown, MA, USA, 2001: 282-289

[32] Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter, Witten Ian H. The WEKA data mining software: an update. SIGKDD Explorations, 2009, 11(1): 10-18

[33] CRF++: Yet Another CRF toolkit. http://crfpp.sourceforge.net/

[34] Pitrelli J F, Beckman M, Hirschberg J. Evaluation of prosodic transcription labeling reliability in the ToBI framework//Proceedings of the International Conference on Spoken Language Processing. Yokahama, Japan, 1994: 123-126

[35] Fry D B. Duration and intensity as physical correlates of linguistic stress. Journal of the Acoustical Society of America, 1955, 27(4): 765-768

[36] Bolinger D L. A theory of pitch accent in English. Word, 1958, 14(2-3): 109-149

[37] Zhao Yuan-Ren. Yu Yan Wen Ti. Beijing: The Commercial Press, 1980(in Chinese)
(赵元任. 语言问题. 北京: 商务印书馆, 1980)

[38] Lin Mao-Can, Yan Jing-Zhu, Sun Guo-Hua. The preliminary experiment of disyllable group normal stress in Beijing dialect. Dialect, 1984, (1): 57-73(in Chinese)
(林茂灿, 颜景助, 孙国华. 北京话两字组正常重音的初步实验. 方言, 1984, (1): 57-73)

[39] Shen Jiong, Hoek J H. The speech mechanism of contrastive accent in Chinese (abstract report). Chinese Research, 1993, (3): 10-15(in Chinese)
(沈炯, Hoek J H. 汉语语势重音的音理(简要报告). 语文研究, 1994, (3): 10-15)

NI Chong-Jia, born in 1979, Ph.D., lecturer. His research interests include speech recognition, prosody modeling, and machine learning.

ZHANG Ai-Ying, born in 1980, M. S., lecturer. Her research interest is machine learning.

LIU Wen-Ju, born in 1960, Ph. D., professor, Ph. D. supervisor. His research interests include speech recognition, speech synthesis, and computational auditory scene analysis.

Background

Prosody is generally used to describe aspects of a spoken utterance's pronunciation which are not adequately explained by segmental acoustic correlates of sound units (phones). The prosodic information associated with a unit of speech, say, syllable, word, phrase, or clause, influences all the segments of the unit in an utterance. They are also referred to as supra-segment that transcends the properties of local phonetic context. In this paper, the main prosodic event that we consider is stress (or prominence, highlighting). Stress refers to the greater perceived strength or emphasis of some syllables in a phrase. Many research have been done in automatic detection of stress (or prominence, highlighting) in speech at both the syllable and word level.

Although English pitch accent detection has been studied extensively, there relatively a few works explore Mandarin stress detection. In this paper, the Mandarin stress detection method is proposed, which is the combination of boosting classification and regression tree (CART) classifier and conditional random fields (CRFs) classifier. Our proposed method can overcome the efficiency of boosting classification and regression tree model, and realize the complementarities with the advantages of boosting classification and regression tree and conditional random fields. The experimental results indicate that our proposed method could acquire better classification effect, and resolve the problem of Mandarin stress detection very well.

This work is supported mainly by the National High Technology Research and Development Program (863 Program) of China Project (Nos. 20060101Z4073, 2006AA01Z194). The main task of this project is to research on the key technology of spontaneous speech to speech translation in the network environment, and to construct a spontaneous speech to speech translation system. The research contents of this project are mainly related to speech recognition, speech understanding, machine translation, speech synthesis, and man-machine interaction. Therefore, a lot of the related technologies can be researched more deeply through the development of this project. Of course, this project has great promise and potential with various applications.

We have made some progresses in this research fields, and some research articles have been published at the international conferences. This research is a part of this project, and especially related to speech recognition. The purpose of this research is to improve the performance of speech recognition system.