文章编号: 1003-0077(2008)04-0088-06

基于韵律信息的连续语流调型评测研究

潘逸倩, 魏 思, 王仁华

(中国科学技术大学 讯飞语音实验室, 安徽 合肥, 230027)

摘 要: 汉语连续语流中的调型评测是汉语语音评测的一个重要环节, 利用连续语流中韵律耦合效应和韵律结构紧密相关这一特性, 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), 使得汉语普通话水平评测系统针对标准连续语流的调型识别率从 82.0% 提升至 84.6%; 针对有方言背景的非标准发音, 机器评分与专家评分的相关度绝对提升超过 3.0%。

关键词: 计算机应用; 中文信息处理; 语音评测; 调型评测; 调型识别; 韵律词; MSD-HMM

中图分类号: TN912.3 **文献标识码**: A

Tone Evaluation of Chinese Continuous Speech Based on Prosodic Words

PAN Yi-qian, WEI Si, WANG Ren-hua

(Man Machine Voice Communication Laboratory, University of Science & Technology of China, Hefei, Anhui 230027, China)

Abstract: The tone evaluation of Chinese continuous speech is a key aspect in Mandarin Chinese pronunciation test. Taking advantage of the close correlation between the prosody framework and the modified tonal curve, this paper presents a Multi-Space Distribution Hidden Markov Model (MSD-HMM) built on the prosodic word for the tone evaluation. The experimental results show that the proposed Mandarin Chinese Pronunciation Evaluation System improves from 82.0% to 84.6% in the performance of tonal syllable error rate for the standard Chinese continuous speech. And for the non-standard Chinese Mandarin speech, the correlation between computer score and expert score achieves over 3.0% absolute improvements compared with that of the baseline system without tone pronunciation test.

Key words: computer application; Chinese information processing; mandarin Chinese pronunciation test; tone evaluation; tone recognition; prosodic word; mandarin speech recognition

1 引言

利用计算机对发音学习者的语音进行标准程度地评估和发音缺陷地检测是计算机辅助语言学习的核心功能。针对汉语这一有调语言, 在语音评测系统中, 调型评测模块是其重要组成部分。

调型识别技术是调型评测的基础。针对发音正确的标准语音, 假设调型识别系统能达到相当高的调型识别率, 则建立在识别技术基础上的调型评测系统就能可靠地评判声调发音质量, 较为准确地分辨出正确和错误的声调。对于孤立字来说, 这一点比较容易实现, 但对于连续语流, 调型识别由于字调受上下文的影响, 存在不可忽视的连续性, 变调现象普遍存在且具有多样性, 从而导致自然语流的调型识别性能常常不佳, 进而影响调型评测系统的实际性能。

基于核心段基频的声调识别方法是处理变调现象的一种方法: Jin-song Zhang[1,2] 提出一个音节的

收稿日期: 2007-12-29 **定稿日期**: 2008-05-06

基金项目: 国家"十五"重点资助项目(ZDI105—B02)

作者简介: 潘逸倩(1982—), 女, 硕士生, 主要研究方向为普通话连续语流调型测试与检错; 魏思(1982—), 男, 博士生, 主要研究方向为计算机辅助语言学习; 王仁华(1943—), 男, 教授, 博导, 主要研究方向为信号与信息处理。

基频可以分成三个部分: 开始段、核心段和结束段,其中最稳定、最具有调型区分性的部分是核心段。例如: 阴平的开始端有时会存在一小段上升曲线,从孤立的角度来判断可能会被误判为阳平;而阳平开始端有时会存在一小段下降曲线,这种曲线的走势模式与上声很相似。作者通过实验证明了基于核心段的声调识别系统相对于基于全音节基频的系统识别率有 6 %的提升。Guokang Fu[3]也提出基于韵母核心段的调型识别方法可以减少数据量,提高调型识别的鲁棒性,同时将该方法运用于其他多种语言上,也得到一定的效果提升。

基于基频核心段的方法,严格地删减每个音节基频段的首尾部分(包括轻声段、噪声段和部分发生变调的基频段),再对核心段两端作延伸补偿,重新生成相对稳定的孤立调型的基频曲线形状,其目的在于去除连续语流中字调之间互相影响最为严重的部分。这种方法的缺点在于核心段的界定在很多情况下比较困难,核心段的完整性、延伸补偿部分的合理性等都尚未找到公认有效的方法。而关于特征内连续性信息的利用,常用的方案是在提取核心段特征的基础上建立声调上下文相关的调型模型,但是这种拓展实质上仍然没有充分利用字调间包含的连续性信息,并且由于没有针对不同韵律结构对不同结构的调型单元分别建模使得模型本身就存在较大的混淆度。

另一种方案是利用将清音部分插值平滑的方法扩展音节基频信息,以两字词或三字词搭配的形式来进行识别,这种方法可以在一定程度上利用字调间的连续特性,但是由于插入的基频段事实上并不存在,从而会增加模型的混淆度,并且在噪声鲁棒性方面也存在较大的问题。

如何合理利用汉语语音声调的超音段信息以及如何解决基频在清浊音之间的不连续性是本文提出并讨论的主要问题。本文以韵律词为模型训练和识别的特征单元,利用多空间状态概率分布建模来处理韵律词单元内清浊音之间的基频连接,得到较高的调型识别率。

本文采用韵律词为调型模型单元,是因为韵律词是最小的语流单位,在韵律词内部没有可感知的停顿,因此在韵律词内部调型之间互相影响的程度比更高层次韵律块之间大,因此存在普遍的调型连变现象。同时加入韵律词单元这一属性,不再使用传统的声调上下文信息,使得同样上下文条件下的边界处与非边界处调型区分建模,更好地模拟了实际发音情况,减小了模型混淆度。

律词调型模型单元由单个或多个音节组成,多个音节间存在清音段,因此如何处理韵律词内部的声调基频特征 F0 的不连续性成为调型曲线建模的难点。我们根据多空间概率分布的隐马尔可夫模型(MSD-HMM)的原理[4],对于基频特征,建立两个概率分布空间,离散的和连续的,分别对应于清音段和浊音段。利用 MSD-HMM 描述基频的观察概率分布可以避免启发式假设和其他人工处理。

本文在第二节描述了基于韵律信息的连续语流调型评测系统原理及系统构成;第三节是实验的结果和分析;在第四节给出结论,并展望今后相关工作的趋势和方向。

2　系统构成

由于本文调型评测系统以调型识别技术为基础,因此如果调型识别系统在发音正确的语音库上识别性能优异,则可以将其技术应用至调型评测系统,并且可以认为该系统对汉语普通话声调发音水平的评价合理可信,因此本文声调评测系统有两个阶段: 标准语音库上的调型识别和方言区语音库上的调型评测。

系统主要包含的五个模块分别为: 基频提取、基频处理、文本分析、模型训练、调型识别/评测。系统框图如下:

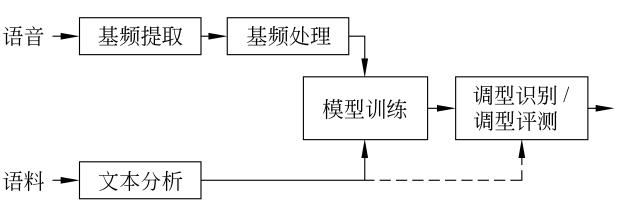


图 1　声调评测系统框图

首先通过基频提取模块从语音中准确而鲁棒提取出基音周期。但是在实际的基音检测中,经常出现半频或倍频的错误估计,并且由于连续语音中声调受到上下文影响,会导致变调现象,与孤立调型单元差距很大。因此第二步基频处理模块的目标是经过一系列的规整处理后得到调型识别所需要的较为准确的基频特征。由于建模单元是韵律词,因此需要将语料通过文本分析模块得到韵律结构标注,从而获得无调音节上韵律词边界。在此基础上利用工具 HTS-2.0,训练得到 16MIX MSD-HMM[4]模型。最后是声调识别/评测模块,采用限定边界的声调识

别方法得到识别率结果和后验概率,再利用谱和基频两部分的后验概率进行线性拟合,得到评测得分。

2.1 特征处理

通过分析不难看出,一段语音的基频特性主要是由以下一些因素所综合决定的: 1) 各孤立音节所对应的纯粹的声调;2) 统计上的一些波动和变化;3) 该段语料文本及其上下文关系;4) 说话人调域。其中 1) 是我们需要得到的有用信息,2) 能被 HMM 本身很好的描述。而 3) 和 4) 则使得声调出现不稳定和变化,从而给声调判决带来困难。

针对上述问题,我们需要对提取的基频进行相应的规整处理。首先使用在 863 连续语音库上训练的 16MIX 的 Mono-Phone 模型对语料进行自动声韵母切分,得到声韵母边界;在此基础上提取韵母核心段,以韵母中间点为中心分别向两端求差分,将大于最大阈值的点定为核心段起始点和结束点;针对不同说话人的影响,还需要对基频作发音人的均值规整;同时为了减少声调上下文之间的相互影响,采用前后各 1 秒的窗来对 F0 进行规整,即长时基音周期规整 (LPN)[5] 算法。文中的基频提取基于 ETSI[6] 工具。

2.2 文本分析

通过上面步骤我们得到较为准确的基音估计,由于我们训练和测试所需要的是以韵律词为单元的特征,因此需要对语料做文本分析,得到相应的韵律结构信息。

由于语音评测系统本身的特殊性,可以首先对训练语料和测试语料做文本分析,主要步骤为文本预处理、分词、分词后处理、声韵母分类,得到每个音节的韵律结构标注。其中对特殊变调要进行纠正,使一些特殊字的声调在组合语音环境中发生规律性的变调进行修正。依据韵律结构标注可以得到训练声调模型所需的韵律边界。

在一般的调型识别系统中由于没有测试文本,可以由韵律词边界所具有的声学特征对声韵母边界(候选边界)分类,生成所需的韵律词边界信息。韵律词边界的声学特征主要表现为: 边界处加入静音段;边界前音节的延长[7];基频重置[8]。

2.3 基于韵律信息的 MSD-HMM 模型

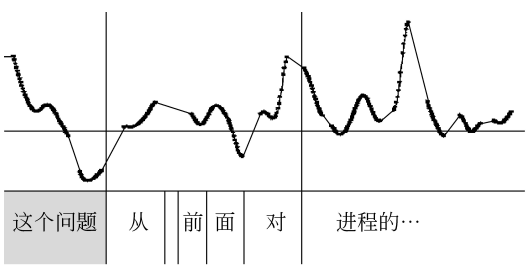
2.3.1 韵律词模型

在语言交流中,韵律信息在语言表达的自然度和可理解程度方面扮演着非常重要的作用[9]。由韵律信息提供的各种层次的间断将连续语流分割成更加便于理解和机器处理的小单元,并且为消除句法歧义提供了重要依据。

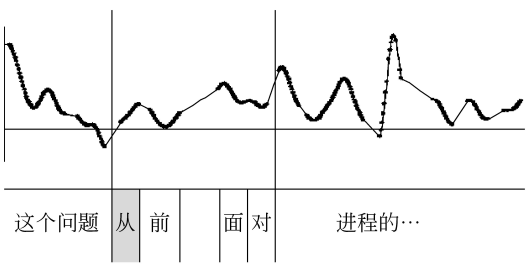
现代语音学研究表明,韵律信息和语流中基频变化密切相关,因此理论上可以根据韵律结构上的基频变化反过来考察调型信息。本文提出的韵律词调型识别系统所需要的特征信息包含两个方面,即基频特征以及与其所对应的韵律层次信息特征。韵律层次信息是描述人们在发音过程中根据语句的语法、语义、语用甚至自身习惯所产生的各种停顿间隔的信息。该信息可以反映出基频的超音段特征的某些规律性。

韵律词是基频变化组,韵律词在由词汇词组成的更大韵律成分中起关键作用。韵律词之间的 F0 重设和韵律词内部各音节 F0 的特有规律性表现是口语的一个重要特点[10]。

我们通过下面的例子将证明韵律词结构对于调型的重要影响。例句如下: "这个问题从前面对进程的论述可以看出"。"从前面对"这个词语组合有两种韵律结构: a. "从前/面对";b. "从/前面/对"。下面利用该句的两种调型曲线对比分析。



(a) "从/前面/对"在连续语流中的基频曲线



(b) "从前/面对"在连续语流中的基频曲线

图 2 "从前面对"在连续语流中的基频曲线

由上图可以看到,基频曲线在连续语流中和孤立词调型差别很大,受上下文影响不可忽视。但是在同样的上下文条件下,发去声的"面"字在(a)中由于在阳平的"前"后面,与其组成一个韵律词,起始端没有达到足够的高度,在(b)中"面"字与后面发去

声的"对"字组成一个韵律词,其结束端没有下降到正常的调值。韵律词内部的制约明显超过更大的韵律块内部的相互影响。在"前"和"对"字上我们也可以看到两种划分结构带来的显著影响。

因此本文采用韵律词为调型模型单元,文中评测系统使用的韵律词边界切分是以文本为基础得到的,该结果是对实际发音中韵律信息的近似处理。真实语音中的停顿、节奏、基频重置情况与每个发音人对于文字的理解、发音习惯都有较强的相关性。为了验证实验中所采用的近似处理方法具有可行性,设计实验如下: 以《普通话水平测试用朗读作品》作品 1 为朗读文本,400 字,共 10 个发音人。该发音数据均为标准数据,发音水平高。以对应的标准语音为分析语料,比较文本分析结果和人工标注的韵律词边界结果。统计结果如下:

表 1　文本分析与人工标注边界对比统计

名　称	定　　义	数目	一致率
TOTAL	文本分析划分的韵律词边界总数	195	-
DIFF1	文本分析与实际发音的不同韵律词边界总数	29	85.1 %
DIFF2	实际发音内部不同韵律词边界数	22	88.7 %

由上述实验可知: 文本分析结果与实际发音的韵律词边界划分结果总的一致率为 85.1 %,10 个发音人之间的韵律词边界一致率为 88.7 %。可以据此分析得出: 1) 发音水平较高的语音数据,韵律词边界信息的一致程度较高;2) 发音水平较高的语音数据,实际韵律词边界信息与文本分析边界划分一致性较高。故我们在实验室中使用文本分析近似代替真实语音韵律层次信息的方法是可行的。

2.3.2　多空间概率分布建模与解码

如何处理声调基频特征 F0 的不连续性是韵律词声调建模的难点。Tokuda 教授等[4] 提出基于多空间概率分布的隐马尔可夫模型(MSD-HMM),用于在语音合成任务中建模基频序列。我们根据 MSD-HMM 的原理,对于基频信息,可以建立两个概率分布空间,离散的和连续的,分别对应于清音段和浊音段。在浊音段区域,基频特征可以看作为由若干个 1 维子空间共同作用产生的观测量,而清音段区域则是由 0 维空间生成。MSD-HMM 可以通过训练调整优化每个状态上各个空间的权值,因此可以通过 MSD-HMM 描述基频的观察概率分布。

假设一个随机事件的观测空间 Ω 由 G 个子独立空间组成,且这些空间的维数可以不同,每个子空间的先验概率为 $p(\Omega_g)$,且 $\sum_{g}^{G} p(\Omega_g) = 1$。观测向量 o,对应到每个子空间有分布概率 $p_g(o)$,因此 o 的观测概率定义为:

$$b(o) = \sum_{g \in S(o)} p(\Omega_g)\, p_g(o) \qquad (1)$$

其中 $S(o)$ 是 o 在子空间的空间序号集,由每个观测值的特征向量决定。

解码过程可以表示为式(2)。

$$\hat{M} = \arg\max_M \prod_t P(q_t \mid q_{t-1}) \times \left[\sum_k c_{kq_t}^p N(o_t^p;\ \mu_{kq_t}^p,\ \sum_{kq_t}^p) \right] \qquad (2)$$

其中 M 代表声调序列,q_t 是 t 时间的状态,而 o_t^p 是基频特征, $\sum_k c_{kq_t}^p N(o_t^p;\ \mu_{kq_t}^p,\ \sum_{kq_t}^p)$ 是基频训练的一个 MSD,$c_{kq_t}^p$ 是混合权重。

2.4　调型评测算法

基于 HMM 的调型识别系统,计算针对指定文本 T 的发音数据 o 的输出概率 $P(T \mid D)$,采用后验概率的方法来进行发音质量评测。连续语流情况下先对连续语流进行切分然后再分段累加。在连续语流中,利用贝叶斯公式,后验概率 $P(T \mid D)$ 计算如(3)式所示。

$$\begin{aligned} P(T \mid O) &= \sum_i^N (|\log(P(T_i \mid O^{(T_i)}))| / NF(T_i)) / N \\ &= \sum_i^N \left(\left| \log \left(\frac{P(O^{(T_i)} \mid T_i)\, P(T_i)}{\sum_{q \in Q} P(O^{(T_i)} \mid q)\, P(q)} \right) \right| / NF(T_i) \right) / N \\ &= \sum_i^N \left(\left| \log \left(\frac{P(O^{(T_i)} \mid T_i)}{\max_{q \in Q} P(O^{(T_i)} \mid q)} \right) \right| / NF(T_i) \right) / N \end{aligned} \qquad (3)$$

其中,Q 是模型集合,q 是 T_i 可能被误读成的音素,$NF(T_i)$ 是音素 T_i 的总帧数,$P(O^{(T_i)} \mid T_i)$ 是由语音识别模型 T_i 得到的针对发音矢量 $O^{(T_i)}$ 的似然度。

以上结果为累加的概率,并不适合直接作为系

统输出打分。需要将累加的后验概率映射成和人工打分可比的评测打分[11]。本系统采用分段线性映射,根据机器预测得到的发音人的对应于普通话水平测试的发音水平等级,将其无调声韵母和调型两部分的后验概率得分转化为输出评测分数。

3 实验结果和分析

3.1 实验配置

数据库文本选自《普通话水平测试用朗读作品》,共有 60 个篇章,每篇文章平均 400 字。训练集为 28 人,包括 15 个女声和 13 个男声的标准语音数据,每人 60 个篇章;测试集为 9 人,每人 60 篇章,包括 4 个女声和 5 个男声的标准语音数据。

3.2 实验结果与分析

实验 1:标准语音库上的调型识别。采用标准语音(普通话水平为一级甲等的语音数据)作为调型识别的测试数据库。此时的调型识别率越高,即表明评测系统性能越可靠。

包含三个子实验,分别为声调上下文相关调型模型(记为 Tritone)的调型识别、基于韵律词并采用插值方法连接基频曲线建模(记为 L1Tone)的调型识别、基于韵律词的 MSD-HMM 模型(记为 L1Tone_MSD)的调型识别。以上实验的特征规整过程包括:提取核心韵母段基频、去除半倍频点、均值规整、LPN 过程(根据实验的具体情况相应过程的处理策略有所不同)。

基线系统(子实验 1):经过特征规整后,训练上下文相关的 16MIX-HMM 模型,其中共有 180 个模型单元。

改进 1(子实验 2):在特征规整处理后,按照文本韵律标注的韵律词结构,对韵律词内清音段部分进行线性插值,以韵律词为模型单位训练 16MIX-HMM 模型,模型数为 123 个单元。包含了 5 种调型的一字、二字、三字韵律组合(其中四字词和五字词已拆开并归入其他类型)。

改进 2(子实验 3):在特征规整处理后,利用工具 HTS-2.0,对标注的韵律词结构对应的音节建立 16MIX MSD-HMM 模型,模型数 123 个单元。

表 2 L1Tone_MSD 中四种调型识别率结果

调型	TONE1	TONE2	TONE3	TONE4
识别率	90.3%	82.6%	84.2%	81.0%

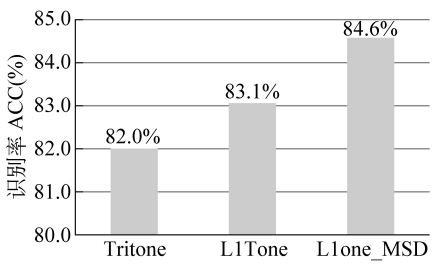


图 3 Tritone\L1Tone\L1Tone_MSD 的识别性能比较

图 3 比较了三种策略下的调型识别结果,其中轻声部分由于具体情况比较复杂没有统计相应的结果。由上图可看出,L1Tone_MSD 的性能最优,L1Tone 次之,Tritone 最差。这是由于上下文相关的声调模型在韵律词边缘处的上下文扩展并不合理,边缘处 Tritone 单元和韵律词内部的情况具有较大差异性,如果作为同一单元的训练样本增加了模型混淆度:L1Tone 方法由于引入插值方法这种人工假设,缺乏真实性,且该问题必然会随着环境信噪比的下降越来越严重,因此性能也不佳。而本文所提出的方法 L1Tone_MSD 可以取得更好的性能。

实验 2:方言区语音库上的调型评测。测试数据为河南、山东地区的带有方言问题的发音数据,100 篇作品,100 人,每篇文章平均 400 字。

本实验中基线系统为建立在河南和山东方言区的汉语普通话发音水平测试系统,主要根据字、词、篇章的频谱特征和字、词的基频特征评测发音人的普通话标准程度[12]。本实验目的是在此基线系统基础上增加基于韵律词的 MSD-HMM 模型的连续语流调型评测子系统(记为 L1Tone_MSD),分析性能是否得到提升。

使用 39 维 MFCC 特征和 3 维基频特征分别训练声韵母模型和声调模型,用两种模型分别对声韵母和调型发音情况进行度量,获得似然度和后验概率等信息,然后用这些信息进行线性拟合,得出机器打分,分析机器打分与专家打分的相关度,结果如图 4 所示。

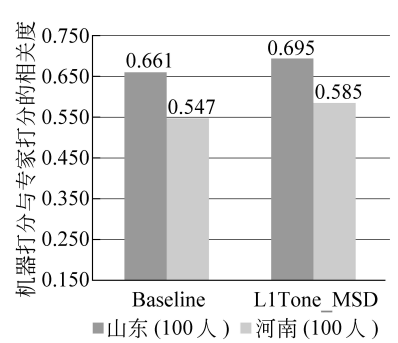


图 4 L1Tone_MSD 与谱特征的解码结果线性拟合后的性能

实验中使用的山东和河南籍发音人的普通话调型错误比较严重，加入调型评测对系统性能的提升有明显的作用。数据显示，在基线系统中加入基于韵律词的 MSD-HMM 模型的调型评测子系统后，机器评分与专家评分的相关度在山东和河南数据上分别提高 3.4 %和 3.8 %。

4　结论和工作展望

实验证明基于韵律信息的多空间状态分布模型，在识别率性能上优于提取核心段后得到的上下文相关模型和基于韵律信息的插值模型。韵律层次信息充分利用了韵律词单元中的调型连续性，并且增加了调型之间的区分性。将基于韵律词的 MSD-HMM 模型的调型评测子系统加入普通话水平评测系统中，使总体评测性能得到了明显地提升。

在今后的工作中，将尝试引入更多韵律信息以提高评测系统的性能。

参考文献：

[1] Kei Kichi Hirose, Jin-song Zhang. Tone Recognition of Chinese Continuous Speech Using Tone Critical Segments[C]// ICSLP98. Sydney, Australia: Dec. 1998, 703-706.

[2] J. Zhang, K. Hirose. Tone Nucleus Modeling for Chinese lexical Tone Recognition[C]// Speech Communication, Vol. 42, 2004, 447-466.

[3] Chen, C.J., Haiping Li, Liqin Shen, Guokang Fu. Recognize Tone Languages Using Pitch Information on the Main Vowel of Each Syllable[C]// 2001, 61-64.

[4] K. Tokuda, T. Masuko, N. Miyazaki, et al. Multi-space Probability Distribution HMM [J]. IEICE TRANSACTIONS on Information and Systems, 2002, E85-D(3): 455-464.

[5] Y. W. Wong, Eric Chang. The Effect of Pitch and Lexical Tone on Different Mandarin Speech Recognition Tasks [C]// EUROSPEECH-2001. 2001, 2741-2744.

[6] European Telecommunications Standards Institute (ETSI) Standard ES 202 050, Extended Advanced Front-end Feature Extraction Algorithm[S].

[7] 林茂灿. 普通话语句的韵律结构和基频(F0)高低线构建[J]. 当代语言学,2002,(4): 254-265.

[8] 冯勇强，初敏，贺琳，吕士楠. 汉语话音节时长统计分析[C]// 第五届全国现代语音学学术会议论文集，2001 年，66-69.

[9] 熊子瑜. 基频重设与语流间断[C]// 第五届全国现代语音学学术会议论文集，2001 年，189-193.

[10] 胡伟湘，徐波，黄泰翼. 汉语韵律边界的声学实验研究[J]. 中文信息学报，2002，16(1): 43-48.

[11] 魏思，刘庆升，胡郁，王仁华. 普通话水平测试电子化系统[J]. 中文信息学报，2006，20(6): 89-96.

[12] SiWei, et al. Putonghua Proficiency Test and Evaluation, Advances in Chinese Spoken Language Processing, Chapter 18 [M]. Springer Press, 2006.