摘要

随着信息全球化发展，人们对文语转换技术的需求也日益迫切，这促使该技术的研究和开发取得了突破性进展。虽然合成语音的质量已经能够使人们感觉基本上可以接受，但其自然度仍然不尽如人意。本文从文语转换系统前端的部分技术——字音转换、韵律短语边界识别、朗读重音判别入手，解决了文语转换中常出现的读音错误，节奏、停顿、轻重方面的处理不当，从而改善汉语语音合成的自然度。

在分析了基于规则和基于统计的方法进行多音字处理的不足之后，提出了较为完备的多音字处理机制：查穷举表、匹配词性规则及计算语义相似度三步骤。本文详细介绍了《知网》，并将其作为语义计算和推演的环境，计算词语相似度及上下文窗口语义相似度。通过实验证明了该多音字处理机制在处理新词和多音词读音方面的优势：语义相似度的计算可以找出未知读音词的语义相似词，通过后者的读音给出前者中多音字的读音。因此该机制不仅使多音字读音判断正确率大幅提高，更改善了文语转换系统的整体字音转换正确率。

虽然一些研究者在韵律短语边界识别方面取得了很好的实验结果，但他们所采用的方法都过于复杂。本文运用在句法分析方面取得成功的基于转换的错误驱动算法建立了一个韵律短语边界识别规则的自学习系统。在建立了一个拥有 5725 句文本和语音相对应的标注语料库之后，通过自学习系统获取了大量规则，这些规则的获取方法不仅简单易行，更在应用于自动识别韵律短语边界时取得了很高的正确率。为了解决大量规则在系统中进行匹配时会带来的负担，本文提出在文语转换系统中树形组织规则，使韵律短语边界自动识别在系统中取得了极佳的实用效果。

尽管汉语是否存在固定的词重音模式还没有定论，但对于文语转换系统来说，汉语重音的研究非常重要。因此，本文从韵律词重音的产生和感知机理入手，对二字词的重音规律进行了分析，并利用同于韵律短语边界识别的自学习系统获取了韵律短语重音规则，但实验结果说明其实用性仍然达不到要求，有待于进一步研究。

关键词：多音字；语义相似度；韵律短语；韵律词；重音

Abstract

Under the development of information globalization, people need text-to-speech technique more and more. This has made outstanding improvement in research and development of text-to-speech technique. Although the quality of synthesized speech has been accepted, the naturalness is still dissatisfying. This paper starts with part of the front end technique, which are text-to-syllable, detection of prosody phrase boundaries and speech stress determining, in order to solve the pronunciation errors and the improper handling in rhythm, pause, stress during text to speech. So the synthesized speech of text-to-speech system can be made more natural.

After analyzing the shortcomings of dealing with polyphones in methods based on rule and on statistics, we introduce a comparatively self-contained flow, which are three steps: look in the polyphone list, fit the part of speech rules and compute semantic similarities. This paper describes How-Net in detail and regards it as the setting of semantic computation and deduction in order to compute the word similarity and the context window similarity. The experiment results show that this kind of method for dealing with polyphones has many virtues: using semantic computing can find the semantic similar word of the unknown pronunciation word, and then get the pronunciation of polyphones in it from its similar word. So this method can improve not only the accuracy of determining polyphone pronunciations greatly but also the accuracy of text-to-syllable in text-to-speech system.

Though some researchers have got some good experiment results in detection of prosody phrase boundaries, their methods are all very complex. In this paper we build a self-learning system for detection rules of prosody phrase boundaries using the transformation-based error-driven learning, which has succeeded in parsing. After building a large corpus that has 5725 text and speech sentences, the self-learning system retrieves a large amount of rules. Not only retrieving of these rules is easy to use but also the accuracy of automatic detection of prosody phrase boundaries can be raised when using these rules. In

order to reduce the burden brought by the large quantity of rules in TTS system, this paper uses tree to organize rules. The automatic detection of prosody phrase boundaries has been used well in real system.

Whether the word stress formulation exists in Chinese mandarin has not got the last word, but research of the Chinese mandarin stress is still very important to text-to-speech system. So this paper starts with the principle of producing and apperceive of prosody word stress and analyzes the rules of two syllable words' stress. Then using the same self-learning system as detection of prosody phrase boundaries to get rules of prosody phrases' stress. But the experiment results are not good enough to be used in the real system. We must research it further in the future.

Keywords polyphone; semantic similarity; prosody phrase; prosody word; stress

目录

第1章 绪论

1.1 研究背景和意义

语音识别和语音合成技术是实现人机语音通信，建立一个有听和讲能力的口语系统所必需的两项关键技术。使电脑具有类似于人一样的说话和听懂人说话的能力，是 90 年代信息产业的重要竞争市场。和语音识别相比，语音合成的技术相对说来要成熟一些，是该领域中近期最有希望产生突破并形成产业化的一项技术[1]。

目前国内外大多数语音合成研究是针对文语转换系统，且只能解决以某种朗读风格将书面语言转换成口语输出，缺乏不同年龄、性别特征及语气、语速的表现，更不用说赋予个人的感情色彩。随着信息社会的需求发展，对人机交互提出了更高的要求，人机口语对话系统的研究也提到了日程上。

提高合成语音的自然度[2]仍然是高性能文语转换的当务之急。就汉语语音合成来说，目前在单字和词组级上，合成语音的可懂度和自然度已基本解决，但是对于句子乃至篇章级，其自然度问题就比较大。

语音合成自然度与人的语言的差距主要体现在两方面：一方面是音质的差距，由于语音合成通常存在一个从语音中提取参数（如音高、音长、音强等），经过适当的变换再生成语音的过程。经过语音到参数再从参数返回到语音的转换过程，恢复出来的语音在音质上往往会有明显的损失，出现杂音、回声、机器声等现象。另一方面是韵律的差距，语音合成系统通常只能生成有限的语调模式，因而使合成语音听起来很单调枯燥。而且语音合成系统还会在节奏、轻重、停顿等方面处理不当，使合成语音听起来很别扭，完整全面的解决，需要自然语言理解的突破。

基于语音数据库的语音合成方法有望进一步提高语音合成的自然度[3]。因为这是一种采用自然语音波形直接拼接的方法，进行拼接的语音单元是从一个预先录下的自然语音数据库中挑选出来的，因此有可能最大限度地保留语音的自然度。但由此产生了一系列新的需要研究的问题，包括：如何确定语音合成的基元，根据什么准则去挑选合适的基元；韵律参数定量化问题[4]，对数据库进行定标问题；以及如何将统计的方法和规则方法相结合使机器能自动发现和找出所需的语音单元，保证最高的合成语句自然度等等。

未来的文语转换系统的发展趋势是采用基于语境相关的合成思想进行设计，能够将发音人的原始发音特征最大限度地保留下来，辅助以先进的层次化语言韵律模型，通过分散统计的模型方法[5]来涵盖语义语音之间的内在联系，使系统能够输出具有高自然度和表现力的合成语音。但是，在目前的合成系统中，普遍存在合成输出语音的机器味比较浓、语境的知识层次模型研究不完善等问题。因此，获得高自然度、具有表现力的合成语音也是今后语音技术的研究目标之一[6]。

实现人机交流是信息时代每个用户的梦想。随着网络的逐渐普及，如何使网站不再冰冷、单调和枯燥，能够“开口”说出人话，使网民上网除了可以“看”网外，还能够做到“听”网，这是每个网站和每个网民的共同愿望。结合网上检索技术，我们可以让网民随心所欲的收听各类信息，正是为了服务于此并结合实验室的项目，在刚刚完成的文语转换系统的基础之上进行改善汉语语音合成自然度方面的研究。

1.2 语音合成技术简介

语音合成或者让计算机说话包含着二个方面的可能性[7]：一是机器能再生一个预先存入的语音信号，就象普通的录音机一样，不同之处只是采用了数字存储技术。简单地将预先存入的单音或词组拼接起来也能作到“机器开口”， 但是“一字一蹦”，机器味十足，人们很难接受。然而如果预先存入足够的语音单元，在合成时采用恰当的技术手段挑选出所需的语音单元拼接起来，也有可能生成高自然度的语句，这就是波形拼接的语音合成方法。为了节省存储容量，在存入机器之前还可以对语音信号先进行数据压缩。另一种可能是采用数字信号处理的方法，将人类发声过程看作是一个模拟声门状态的源，去激励一个表征声道谐振特性的时变数字滤波器，这个源可能是周期脉冲序列，它代表浊音情况下的声带振动，或者是随机噪声序列，代表不出声的清音。 调整滤波器的参数等效于改变口腔及声道形状，达到控制发不同音的目的，而调整激励源脉冲序列的周期或强度，将改变合成语音的音调、重音等。因此，只要正确控制激励源和滤波器参数（一般每隔 10～30ms送一组），这个模型就能灵活地合成出各种语句来，因此又称作为参数合成的方法。根据时变滤波器的结构形式不同，又有LPC合成和共振峰合成器等之分。

按照人类言语功能的不同层次，语音合成也可分成三个层次，它们是：

（1）从文字到语音的合成（Text-To-Speech）；（2）从概念到语音的合成（Concept-To-Speech）；（3）从意向到语音的合成（Intention-To-Speech）。这三个层次反映了人类大脑中形成说话内容的不同过程，涉及人类大脑的高级神经活动。不难想象，即使是按规则的文字到语音合成（文语合成）也已经是相当困难的任务。为了合成出高质量的语言，除了依赖于各种规则，包括语义学规则、词汇规则、语音学规则外，还必须对文字的内容有很好的理解，这将涉及自然语言理解的问题。从这一点讲，文语转换系统（以下简称 TTS）实际上也可看作一个人工智能系统。图 1-1 显示了一个完整的文语转换系统示意图。文语转换过程是先将文字序列转换成音韵序列，再由语音合成器生成语音波形。其中第一步涉及语言学处理，例如分词、字音转换等，以及一整套有效的韵律控制规则[8]；第二步需要先进的语音合成技术，能按要求实时合成出高质量的语音流。因此一般说来，文语转换系统都需要一套复杂的文字序列到音素序列的转换程序，也就是说，文语转换系统不仅要应用数字信号处理技术，而且必须有大量的语言学知识的支持。当然其中语音合成终究还是最基本的部分，它相当于“人工嘴巴”，任何语音合成系统包括文语转换系统，都离不开语音合成器。

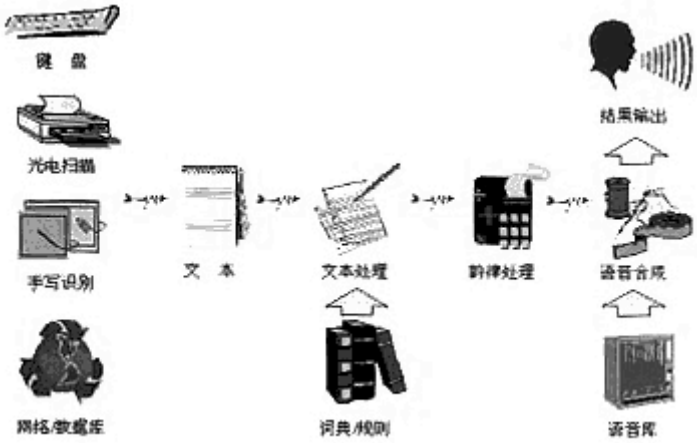


图 1-1 文语转换系统示意图

Figure 1-1 Text-to-Speech System Sketch Map

1.3 国内外语音合成技术发展现状

综观语音合成技术的研究已有二百多年的历史，但是真正有实用意义的近代语音合成技术是随着计算机技术和数字信号处理技术的发展而发展起来

的，主要是让计算机能够产生高清晰度、高自然度的连续语音。近几十年来国际和国内的研究主要集中在按规则文语转换，即将书面语言转换成口头语言。在语音合成技术的发展中，早期的研究主要是采用参数合成方法。值得提及的是Holmes的并联共振峰合成器（1973）和Klatt的串/并联共振峰合成器（1980），只要精心调整参数，这两个合成器都能合成出非常自然的语音。而最具代表性的文语转换系统数美国DEC 公司的DECtalk（1987），该系统采用Klatt的串/并联共振峰合成器，可以通过标准的接口和计算机连网或单独接到电话网上提供各种语音信息服务，它的发音清晰，并可产生七种不同音色的声音，供用户选择。但是经过多年的研究与实践表明，由于准确提取共振峰参数比较困难，虽然利用共振峰合成器可以得到许多逼真的合成语音，但是整体合成语音的音质难以达到文语转换系统的实用要求。自八十年代末期至今，语音合成技术又有了新的进展，特别是基音同步叠加（PSOLA）方法的提出（1990），使基于时域波形拼接方法合成的语音的音色和自然度大大提高。九十年代初，基于PSOLA技术的法语、德语、英语、日语等语种的文语转换系统都已经研制成功。这些系统的自然度比以前基于LPC方法或共振峰合成器的文语转换系统的自然度要高，并且基于PSOLA方法的合成器结构简单易于实时实现，有很大的商用前景。最近几年，一种新的基于数据库的语音合成方法正引起人们的注意。在这个方法中，合成语句的语音单元是从一个预先录下的庞大的语音数据库中挑选出来的，不难想象只要语音数据库足够大，包括了各种可能语境下的语音单元，理论上讲有可能拼接出任何语句。由于合成的语音基元都是来自自然的原始发音，合成语句的清晰度和自然度都将会非常高[9]。

国内的汉语语音合成研究起步较晚些，但从八十年代初就基本上与国际上研究同步发展。大致也经历了共振峰合成、LPC合成至应用PSOLA技术的过程。在国家 863 计划，国家自然科学基金委，国家攻关计划，中国科学院有关项目等支持下，汉语文语转换系统研究近年来取得了令人举目的进展，其中不乏成功的例子：如中国科学院声学所的KX-PSOLA（1993），联想佳音（1995）；清华大学的TH_SPEECH(1993)；中国科技大学的KDTALK（1995）等系统。这些系统基本上都是采用基于PSOLA方法的时域波形拼接技术，其合成汉语普通话的可懂度、清晰度达到了很高的水平。然而同国外其它语种的文语转换系统一样，这些系统合成的句子及篇章语音机器味较浓，其自然度还不能达到用户可广泛接受的程度，从而制约了这项技术的大规模进入市场[10]。

1996 年哈尔滨工业大学机器翻译实验室在达雅翻译工作站中加入了达雅朗读子系统，这是哈尔滨工业大学首次介入汉语语音合成研究领域[11]。2000 年又重新开始这方面的研究，并于 2001 年开发出新的TTS演示系统。

1.4 实验系统简介

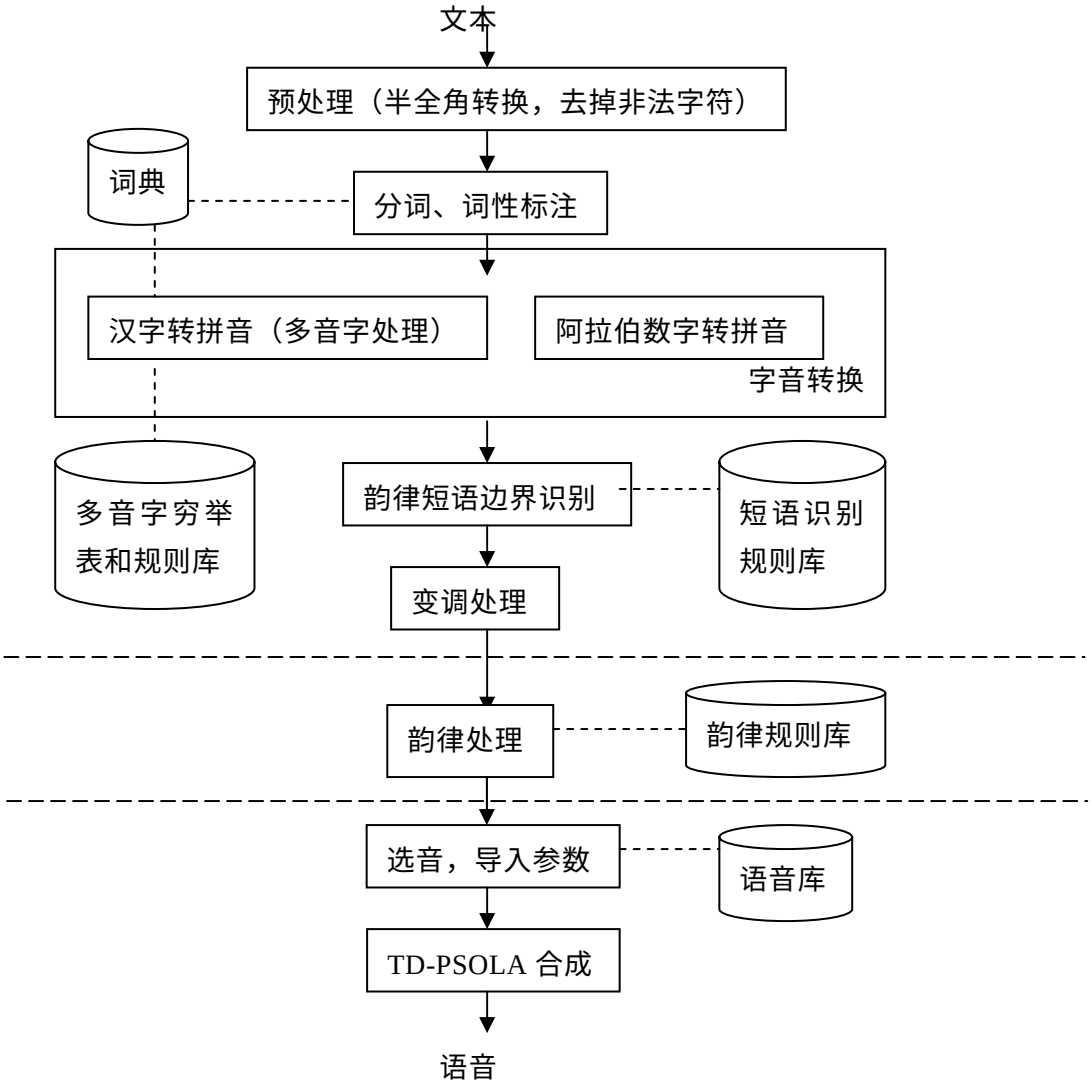


图 1-2 TTS 演示系统结构图

Figure 1-2 the Frame of TTS Demo System

2001 年作者与所在实验室的一名本科生在仅拥有少量资源的基础上搭建了一个汉语文语转换演示系统。这个演示系统采用的语音合成技术是当前

广泛使用的 PSOLA 算法。整个演示系统的构成如图 1-2。

文语转换系统主要由三个部分组成：

1) 语言学处理：在文语转换系统中起着重要的作用，主要模拟人对自然语言的理解过程——文本规整、词的切分、语法分析和语义分析，使计算机对输入的文本能完全理解，并给出后两部分所需要的各种发音提示。
2) 韵律处理：为合成语音规划出音段特征，如音高、音长和音强等，使合成语音能正确表达语意，听起来更加自然。
3) 声学处理：根据前两部分处理结果的要求输出语音，即合成语音。

由于演示系统的模块清楚独立，易于对各个模块的实验结果进行评价，因此这里以该演示系统作为本文的实验系统。

1.5 本文主要研究内容和组织

为了合成出高质量的语音，除了依赖于各种规则，包括语义学规则、词汇规则、语音学规则外，还必须对文字的内容有很好的理解，这将涉及自然语言理解的问题。文语转换过程是先将文字序列转换成音韵序列，再由语音合成器生成语音波形。其中第一步涉及语言学处理，例如分词、字音转换等，以及一整套有效的韵律控制规则；第二步需要先进的语音合成技术，能按要求实时合成出高质量的语音流。因此一般说来，文语合成系统都需要一套复杂的文字序列到音素序列的转换程序，也就是说，文语转换系统不仅要应用数字信号处理技术，而且必须有大量的语言学知识的支持。因此本文下面从这里入手先解决读音错误问题，再对改善合成语音自然度进行研究：

第二章介绍《知网》的原理和构造以及多音字的处理技术，提出了一种利用《知网》通过语义相似度计算判定多音字读音的方法，并根据实验结果证明了此方法的优越性。

第三章描述了实验语料库建设，阐述了基于转换的错误驱动算法自动获取韵律短语边界的改进算法，并提出了在系统中以树形结构组织规则来提高 TTS 运行效率。

第四章从理论上分析了朗读重音在 TTS 中的作用与表现形式，利用一些数据对此进行了说明，并进行了韵律短语重音规则自动获取实验。

第2章 语义处理多音字

2.1 多音字处理

一般来说，每个汉字都有其固定的发音，有些特殊的汉字有两个或多个发音[12]。人可以根据自己的语言经验来确定其发音，然而TTS系统缺乏语言学知识，不能完全准确地进行字音转换。一字多音是汉语中常见的语言现象，前 10 个高频字中就有 6 个是多音字（的、一、了、不、和、大），所以正确地处理多音字不仅保证语音输出的正确性，而且可以改善文语转换的可懂度和自然度。一字多音一般可分为两种情况：一种是由于语音流变现象造成的，这种情况可借助语言学知识建立相应的变调规则进行变音处理。第二种是指汉字本身是多音字[13]。

在 TTS 中进行字音转换时，通常的方法是，对于非单字词直接从词典中提取出该词的读音；对于单字词看是否多音字，若不是多音字则直接给出其读音，若是多音字，则给出该多音字的高频音。这种方法虽然有时也能得到满意的结果，但整体注音准确率最多 85%，而且对于多音词的读音问题不能有效地进行处理。表 2-1 给出了多音词的例子。

表 2-1 多音词举例

Table 2-1 Examples of Polyphone Word

例词	读音	例句
好事	hao3 shi4	这对所有人来说都是件好事
	hao4 shi4	他是个好事之徒
澄清	cheng2 qing1	你必须把事实澄清
	deng4 qing1	不把这桶水澄清不能拿去用
转向	zhuan3 xiang4	他把头转向了右侧
	zhuan4 xiang4	她一到这栋楼就有点转向

通过对现有的 TTS 文本语料库中多音字词的分析我们可以看出，这些多音字词在读音不同时它的词性以及上下文都是不同的，所以可以编写一些规则来更好的处理这些多音字词。这也就是当前汉语 TTS 中最常用的多音字处理方法——基于规则的字音转换方法。但这种方法很难适应大规模开放性的语料，因为自然语言是非常复杂的，仅仅一些规则很难将自然语言中的各种多音字词现象完全描述清楚。对于一些规则库中没有穷举到的词（本文

称其为“新词”），规则的方法更是无法解决其读音的判定问题。另外，规则的编写不仅需要投入大量的人力、物力，更需要编写者有着很深的语言学功底，因而规则方法有着很大的局限性。

随着统计方法在语音识别技术的广泛应用，越来越多的人将统计方法应用于自然语言处理领域，也有人提出基于统计的字音转换。由于多音字要远比同音字的数量少，而且只计算多音字词同其他词的同现，所以参数空间较语音识别有所下降。这样在空间和时间上统计方法都能够满足 TTS 字转拼音的要求，并能很好地解决一些字音转换出现的问题。但是统计方法同样有局限性，它对统计语料有很大的依赖性，它要求统计语料要非常全面，并且语料需要人工校对，工作量仍然非常巨大。而训练语料的不足对字转拼音的正确率会有很大影响。

进而有人提出将规则的方法与统计的方法结合起来，通过实验证明取得了很好的效果。但易见仍无法摆脱语料质量的束缚。

众所周知，一个字的读音与其字义相关，读音不同表达的意思也不同，反过来说，不同的意思使该字的读音不同。也正是这个原因如果TTS输出的读音错误将导致听众对原文的误解或不解。如果能运用语义来进行多音字处理，则直接利用了多音字的根本。近几年来，一些大规模、可计算的语义知识库，包括WordNet、MindNet、FrameNet等的开发和利用，为进行大规模的真实文本的语义分析和理解提供了有利的支持[14]。而《同义词词林》和《知网》的出现更为国内进行汉语语义分析提供了资源上的支持[15]。考虑到汉语语义与读音的关联，下面我们利用《知网》来进行多音字处理，尤其是新词与多音词读音的判定。

2.2 《知网》

在引出利用《知网》处理多音字的方法之前，先介绍一下知网。《知网》（英文名称 HowNet）是其创建人董振东先生花费逾十年研究心血的重要成果。《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，它是一个网状的有机的知识系统[16]。

2.2.1 《知网》的哲学

知网系统的哲学也就是它对客观世界的认识与把握。知网哲学的根本点

是：世界上一切事物（物质的和精神的）都在特定的时间和空间内不停地运动和变化。它们通常是从一种状态变化到另一种状态，并通常由其属性值的改变来体现。试以人为例，人的生老病死是一生的主要状态。这个人的年龄（属性）一年比一年大｛属性值｝，随着年龄的增长头发的颜色（属性）变为灰白｛属性值｝。另一方面，一个人随着年龄的增长他的性格（精神）变得日益成熟｛属性值｝，他的知识（精神产品）愈益丰富｛属性值｝。基于上述，知网的运算和描述的基本单位是：万物，其中包括物质的和精神的两类，部件，属性，时间，空间，属性值以及事件。

2.2.2 《知网》的特点

计算机化是知网的重要特色。知网是面向计算机的，是借助于计算机建立的，将来可能是计算机的智能构件。

知网作为一个知识系统，是一个网而不是树。它所着力要反映的是概念的共性和个性，例如：对于“医生”和“患者”，“人”是它们的共性。知网在主要特性文件中描述了“人”所具有的共性，那么“医生”的个性是他是“医治”的施事，而“患者”的个性是他是“患病”的经验者。对于“富翁”和“穷人”，“美女”和“丑八怪”而言，“人”是它们的共性。而它们的个性，即：“贫”、“富”与“美”、“丑”等不同的属性值，则是它们的个性。

同时知网还着力要反映概念之间和概念的属性之间的各种关系。知网把一种知识网络体系（如图 2-1）明确的教给了计算机进而使知识对计算机而言是可操作的。

总的来说，知网描述了下列各种关系：

a) 上下位关系
b) 同义关系
c) 反义关系
d) 对义关系
e) 部件-整体关系（由在整体前标注 % 体现，如“心”,“CPU”等）
f) 属性-宿主关系（由在宿主前标注 & 体现，如“颜色”,“速度”等）
g) 材料-成品关系（由在成品前标注 ? 体现，如“布”,“面粉”等）
h) 施事/经验者/关系主体-事件关系（由在事件前标注 * 体现，如“医生”,“雇主”等）

i) 受事/内容/领属物等-事件关系（由在事件前标注 $ 体现，如“患者”,“雇员”等）
j) 工具-事件关系（由在事件前标注 * 体现，如“手表”,“计算机”等）
k) 场所-事件关系（由在事件前标注 @ 体现，如“银行”,“医院”等）
l) 时间-事件关系（由在事件前标注 @ 体现，如“假日”,“孕期”等）
m) 值-属性关系（直接标注无须借助标识符，如“蓝”,“慢”等）
n) 实体-值关系（直接标注无须借助标识符，如“矮子”,“傻瓜”等）
o) 事件-角色关系（由加角色名体现，如“购物”,“盗墓”等）
p) 相关关系（由在相关概念前标注 # 体现，如“谷物”,“煤田”等）

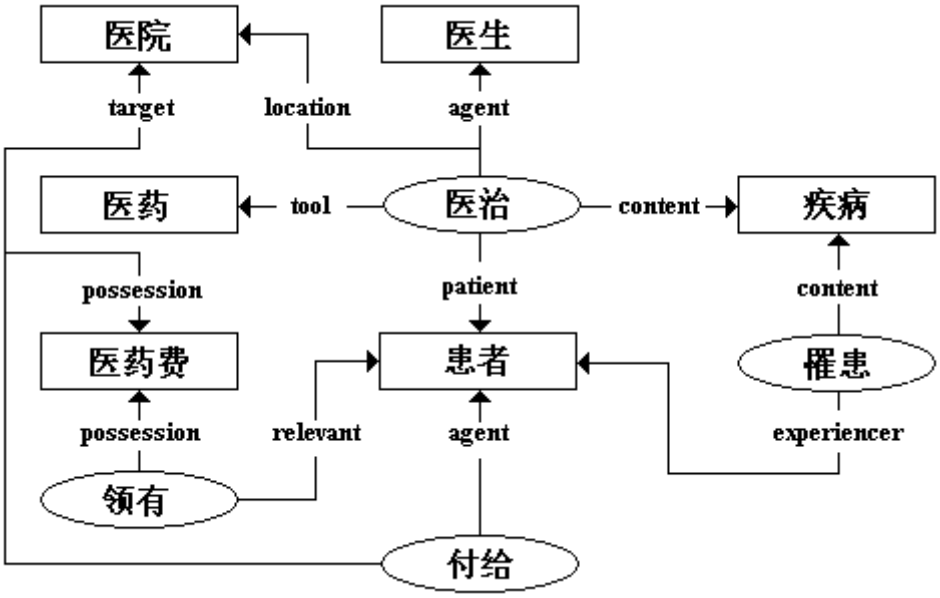


图 2-1 知网关系举例

Figure 2-1 Example of How-Net Relation

知网是一个知识系统，而不是一部语义词典。尽管被我们称为知识词典的常识性知识库是知网的最基本的数据库。知网的全部的主要文件包括知识词典构成了一个有机结合的知识系统。例如，主要特征文件、次要特征文

件、同义、反义以及对义组的形成，以及事件关系和角色转换等都是系统的重要组成部分，而不仅仅是标注的规格文件。

2.2.3 知网的基本单位——义原

义原是最基本的、不易于再分割的意义的最小单位。例如：“人”虽然是一个非常复杂的概念，它可以是多种属性的集合体，但我们也可以把它看作为一个义原。设想所有的概念都可以分解成各种各样的义原，同时应该有一个有限的义原集合，其中的义原组合成一个无限的概念集合。如果能够把握这一有限的义原集合，并利用它来描述概念之间的关系以及属性与属性之间的关系，就有可能建立所设想的知识系统。在《知网》中，利用汉字进行考察和分析来提取这个有限的义原集合。因为汉语中的字（包括单纯词）是有限的，并且它可以被用来表达各种各样的单纯的或复杂的概念，以及表达概念与概念之间、概念的属性与属性之间的关系。义原的提取过程是这样的：对大约六千个汉字列出其全部义原，重复的加以合并，最后整理得到 1503 个义原。

这 1503 个义原的分布如下：

实体——1 个

- 万物（物质、精神、事情、组织）——134 个
- 部分（部件、配件）——3 个
- 时间——1 个
- 空间（方向、位置）——3 个
- 事件（关系/状态、动作）——813 个
- 属性值（外观、量度、特性、关系、状况）——316 个
- 数量值——13 个

属性——117 个

数量——3 个

专项特征（如“领域”等）——99 个

2.2.4 知网系统的概貌

知网系统包括的数据文件和程序有知网管理系统、中英双语知识词典。《知网》的规模主要取决于双语知识词典数据文件的大小。由于它是在线的，修改和增删都很方便，因此它的规模是动态的。它的规模通常以词语的

条数以及由词语所表述的概念的条数计算。

知识词典是《知网》系统的基础文件。在这个文件中每一个词语的概念及其描述形成一个记录。每一种语言的每一个记录都主要包含 4 项内容。其中每一项都由两部分组成，中间以“ = ”分隔。每一个“ = ”的左侧是数据的域名，右侧是数据的值。它们排列如下：

W_X= 词语

E_X= 词语例子

G_X= 词语词性

DEF= 概念定义

其中的 W_X、G_X、E_X 构成每种语言的记录，X 用以描述记录所代表的语种，X 为 C 则为汉语，为 E 则为英语。每个词语由 DEF 来描述其概念定义，DEF 的值由若干个义原及它们与主干词之间的语义关系描述组成。

概念定义是最重要的部分，《知网》是通过它来体现概念与概念和概念的属性与属性之间的相互关系，因此，《知网》对于概念的描述必然是复杂的。这就必须有一套明确的规范，否则便无法保证描述的复杂度和描述的一致性。概念描述既有总的、一般性的描述，也有因不同类别的细节性描述。

概念定义描述的总规定：

(1) 任何一个概念的 DEF 项是必须填写的，不得为空。

(2) DEF 项中用以定义的特性至少是一个，但也可以是多个，数量没有限制，只要内容是合理的且形式是合乎规范的。

(3) DEF 项的第一位置所标注的必须是知网文件所规定的主要特征，否则视为语法错误。但是有些关系意义，可以把次要特征置于{}中后，作为第一位置标注。例如一些介词、连词等虚词，严格地说它们本身没有概念意义。

(4) 多个特征之间应以英文逗号“ ,”分隔，且逗号与特征之间没有空格。

(5) 除第一位置以外，其他位置也可以填有主要特征，但应该说明的是，当主要特征在非第一位置时它失去了原有的上下位关系。

(6) DEF 项中任何一个位置上的信息都可以带有知网所规定的标示符号。

从规定中可以了解到，概念定义的第一项为词的上位关系，而其它项则定义了《知网》所描述的其它关系，其它项之间也可能彼此存在着某些关系。

《知网》的知识词典是以词语及其概念为基础的。

关于词语的例子：

NO.=017140

W_C=打

G_C=V

E_C=~网球，~牌，~秋千，~太极，球~得很棒

W_E=play

G_E=V

E_E=

DEF=exercise|锻练,sport|体育

上例中 E_C 项的“ ~ ”号代表 W_C 项的词。通过 DEF 的定义我们可以知道在“ 打球 ”中“ 打 ”和“ 体育 ”与“ 锻练 ”有关。

除了语义词典外，知网还提供了义原分类树，分类树把各个义原及它们之间的联系以树的形式组织在一起，父子结点的义原具有上下位的关系。我们可以通过义原分类树计算义原间的语义距离。在知网中存在 ENTITY、

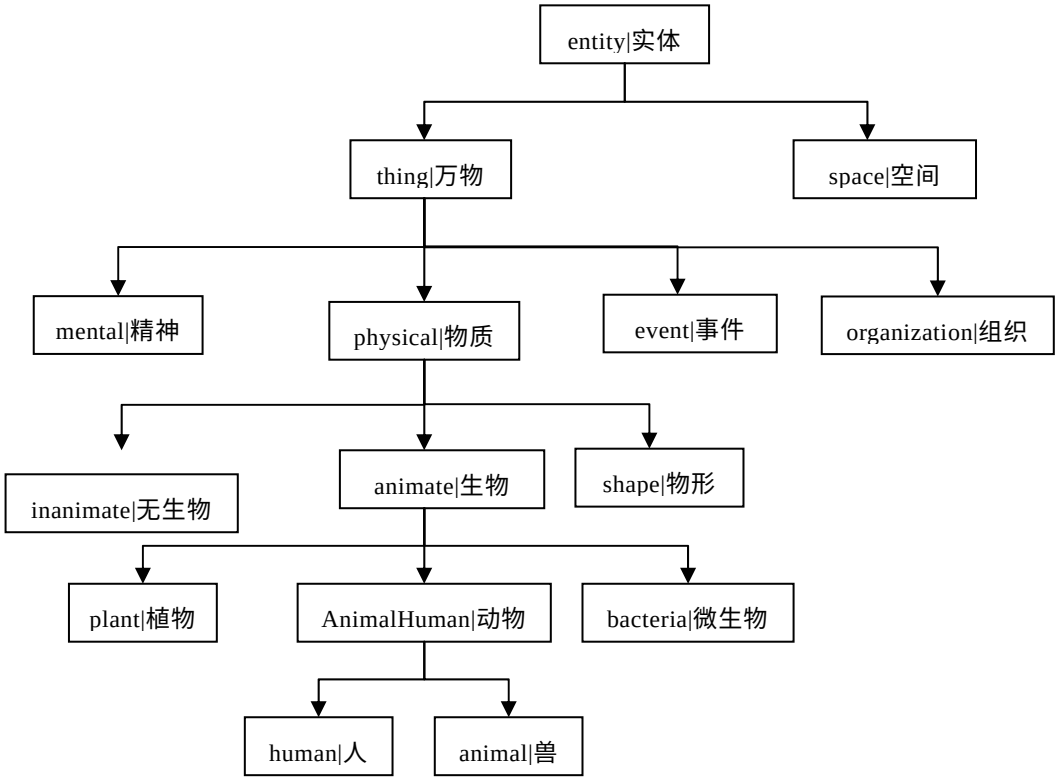


图 2-2 ENTITY 义原分类树的树结构表示

Figure 2-2 the Frame of “ENTITY” Sememe Category Tree

EVENT 等几棵分类树，如图 2-2 是表示实体义原关系的 ENTITY 分类树。

在上面这株义原分类树中“ entity|实体 ”是最高层的概念，其子节点“ thing|万物 ”和“ space|空间 ”是较之低一层的概念。下面的层次依次类推，比如“ human|人 ”属于“ entity|实体 ”但是比“ entity|实体 ”低 5 级的概念。

2.3 利用《知网》处理多音字

既然《知网》是可根据义原进行计算的，那么就在考虑多音字词上下文的前提下在《知网》中找出已经确定读音的字词，从而得到该多音字词的读音。

2.3.1 对《知网》的改造

我们这里的多音字处理研究是针对汉语的，于是把《知网》中的英语词条都去掉，并去掉相应的英语义原标识，这样剩下 6 万多汉语词条。另外，在《知网》的语义词典中没有标注读音，因此需要在每个词条中加入一项“ Y_C ”用来引入该词条相应的拼音。最后得到经过改造的《知网》的词条含有 5 个义项[17]，例如：

W_C=扒

G_C=V

E_C=~猪肉，鸡爪~豆腐，~鸡

Y_C=pa2

DEF=烹调

“ pa ”为拼音音节，“ 2 ”为声调，表示阳平，这样合起来“ pa2 ”标识“ 扒 ”字在这个含义下的读音。我们用 1 至 5 的数字表示声调的阴平、阳平、上声、去声和轻声。

2.3.2 语义相似度

从上一节讲到的义原分类树可以看出，任何在同一株义原分类树中的义原都可以计算它们在这株树中的距离。而树中的层次代表义原的语义位置，因此可以计算任何两个义原的语义相似度。每个词条的概念定义都是由若干义原组合而成的，于是词条之间的语义相似度也可以通过义原的语义相似度

得出。另外，有时还需要进行考虑上下文前提下的语义相似度计算，于是增加一个观察窗口，根据窗口内所有词语的相似度来计算窗口的语义相似度。下面给出义原相似度、词语相似度和上下文窗口相似度的定义：

2.3.2.1 义原相似度[18]

定义：义原 a 与 b 的语义距离

DISTANCE-ATOM(a,b)=a 与 b 在义原分类树上的最短距离

举例：

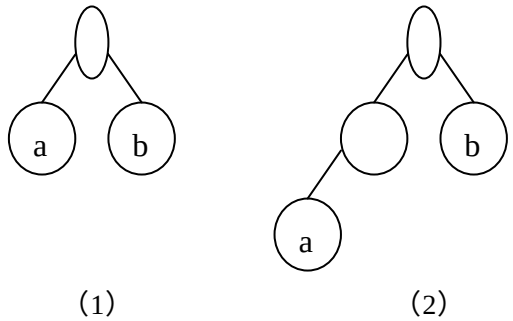


图 2-3 义原距离示意图

Figure 2-3 Sketch Map of the Distance of Sememes

图 2-3 中（1）所表示的情况下 DISTANCE-ATOM(a,b)=2；

图 2-3 中（2）所表示的情况下 DISTANCE-ATOM(a,b)=3；

定义：义原 a 与 b 的语义相似度

SIM-ATOM(a,b)＝

(1－DISTANCE-ATOM(a,b)/ (义原分类树树高×2)) ×100 (2-1)

2.3.2.2 词语相似度

定义：设单词义项 I 概念定义中的所有义原的前缀语义关系的集合为 Relations(I)，其中所有前缀语义关系为 K 的义原集合为 Item-Relation-Atoms(I,K)，则单词义项 I，J 的相似度为：

$$\text{SIM-WORD-ITEM(I, J)}=\frac{\sum_{\text{K}\in\text{Relation(I)}}\ \max_{\substack{\text{a}\in\text{Item-Relation-Atom(I,K)}\\ \text{b}\in\text{Item-Relation-Atom(J,K)}}} SIM-ATOM(a,b)}{\left|\text{Re}lation(I)\right|} \qquad (2\text{-}2)$$

上式中$\left|\text{Re}lation(I)\right|$为单词义项 I 概念定义中的所有义原的前缀语义关系的个数。

2.3.2.3 上下文窗口相似度

定义：上下文窗口的观察长度为窗口内词语个数，设窗口 A 中的所有词语集合为 Item-Word(A)，窗口 A,B 语义相似度为：

SIM-WINDOW-ITEM(A,B)=

$$\frac{\sum_{X \in Item-Word(A)} \underset{Y \in Item-Word(B)}{Max} SIM-WORD-ITEM(X,Y)}{|Item-Word(A)|} \tag{2-3}$$

上式中$|Item-Word(A)|$为窗口 A 的长度。

2.3.3 多音字处理算法

考虑到《知网》的语义词典和义原分类树的规模庞大，完全依靠其进行 TTS 中多音字字转拼音会使系统效率低下，于是仅在遇到新词和多音词时再进入语义处理多音字模块。本文所用实验系统的字转拼音模块流程如下：

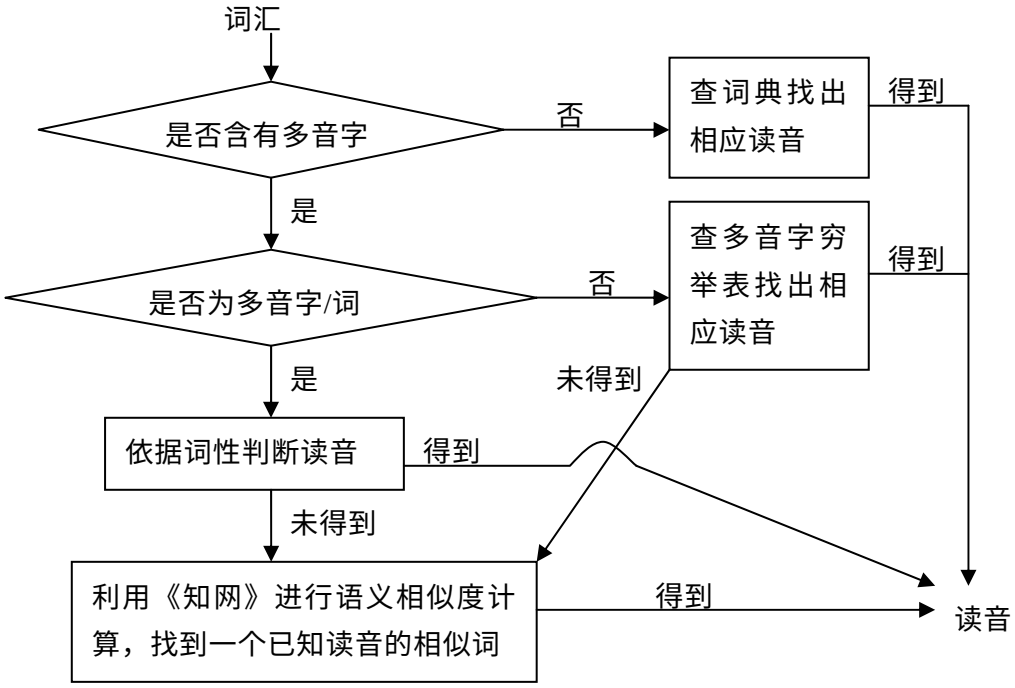


图 2-4 实验系统字转拼音流程

Figure 2-4 Flow of Text to Syllable of Experiment System

由于语义计算能够在相似的词语间对读音进行推演，因而对多音字的穷举表要求并不高，只要穷举出所有多音字在某种语义类别下的少数几个实例

就行了。而《知网》语义词典中的拼音项“ Y_C ”绝大多数也可以为空，仅把常用的四百多个多音字的相应词条对应该语义下的读音标出即可。这样利用《知网》进行处理的词汇或者是多音词或者是没有穷举到的新词。

下面详细说明利用《知网》计算语义相似度进行多音字读音判别的算法：

在进行多音字处理之前已经进行了分词、词性标注，因而得到的输入形式为一个四元组 I=(W,P,T,X)。W 是多音字所在词语，P 为多音字，T 为多音字所在词语的词性，X 为多音词标志（值 1 则是多音词，值 0 则非多音词）。

第一步：在《知网》语义词典中查找所有 W_C=W 且 G_C＝T 的词条，找不到转第二步，找到则看该词条的 Y_C 是否为空。为空且 X=0，转第二步；不为空且 X=0 说明 Y_C 标识的读音即为该多音字在这个词语中的读音，输出；X=1 则转第四步。

第二步：在《知网》语义词典中查找所有 W_C=P 且 G_C＝T 的词条，看是否某一个词条的 E_C 中含有 W，则这个词条的 Y_C 标识的读音即为该多音字在这个词语中的读音，输出；否则，转第三步。

第三步：计算 W 与第二步搜集到的所有 W_C=P 且 G_C＝T 的词条的 W_C 计算词语相似度，取相似度最大的词条中 Y_C 标识的读音作为输出。

第四步：提取 W 在句子中的前一个词语和后一个词语与 W 一起形成一个长度为 3 的上下文窗口，把《知网》语义词典中 W_C=W 且 G_C＝T 的每一个词条单独作为一个长度为 1 的上下文窗口。长度为 3 的窗口分别与长度为 1 的进行相似度计算，取相似度最大的对应的窗口长度为 1 的词条中的 Y_C 标识的读音输出。

2.3.4 举例说明多音字处理算法

先举例说明多音词字转拼音过程。句子“ 不把这桶水澄清不能拿去用 ”分词、词性标注后的结果为“ 不/d把/p这/r 桶/ng 水/ng澄清/vg不/d 能/vz拿去/vg 用/p ”，其中“ 澄清 ”为多音词，可能的发音为“ cheng2 qing1 ”和“ deng4 qing1 ”。由于词性标注[19]所用的标注集不同于《知网》的词性标注集，因此这里先将词性标注符号映射成《知网》的词性标注符号，映射表见附录表A-1。词性标注符“ vg ”映射成“ V ”，则输入的四元组为（澄清,澄,V,1）。运用语义处理算法的流程如图 2-5。

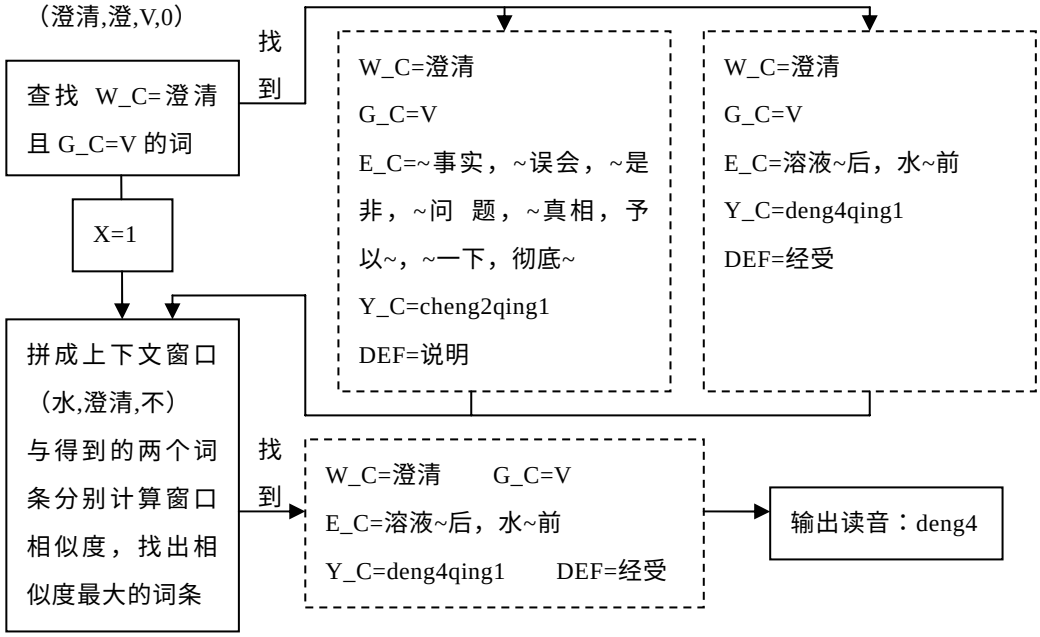


图 2-5 “ 澄清 ” 中 “ 澄 ” 的读音判断过程

Figure 2-5 Syllable Determining Flow of “澄” in “澄清”

再举一个例子说明新词中多音字字转拼音的过程。句子 “ 他今天去银行取钱 ” 分词、词性标注后的结果为 “ 他/r 今天/t 去/vq 银行/ng 取/vg 钱/ng ”，其中 “ 银行 ” 含有多音字 “ 行 ”。词性标注符 “ ng ” 映射成为 “ N ”，则输入的四元组为（银行,行,N,0）。运用语义处理算法的流程如图 2-6。

2.4 实验结果与分析

为了测试语义处理多音字算法进行多音词和新词字转拼音的效果，我们做了一些实验。在做实验之前首先要准备测试语料。

在我们用于韵律短语分析的语料中有文本 10,000 多句，标注好的每句的拼音都已经与相应的文本进行了对齐并作了分词、词性标注。统计发现，常用的 369 个多音字共在 6,000 多个不同的词里出现了 40,000 多次。于是我们取出这些词以及其上下文（前一个词、后一个词）进行脱离词典及穷举表的实验，也就是说把所有这些词都假定为新词或多音词。

我们把 369 个多音字按音序排列（详见附录表 A-2），取出前 40 个进行测试。这些多音字的字转拼音正确率为 91.4%。

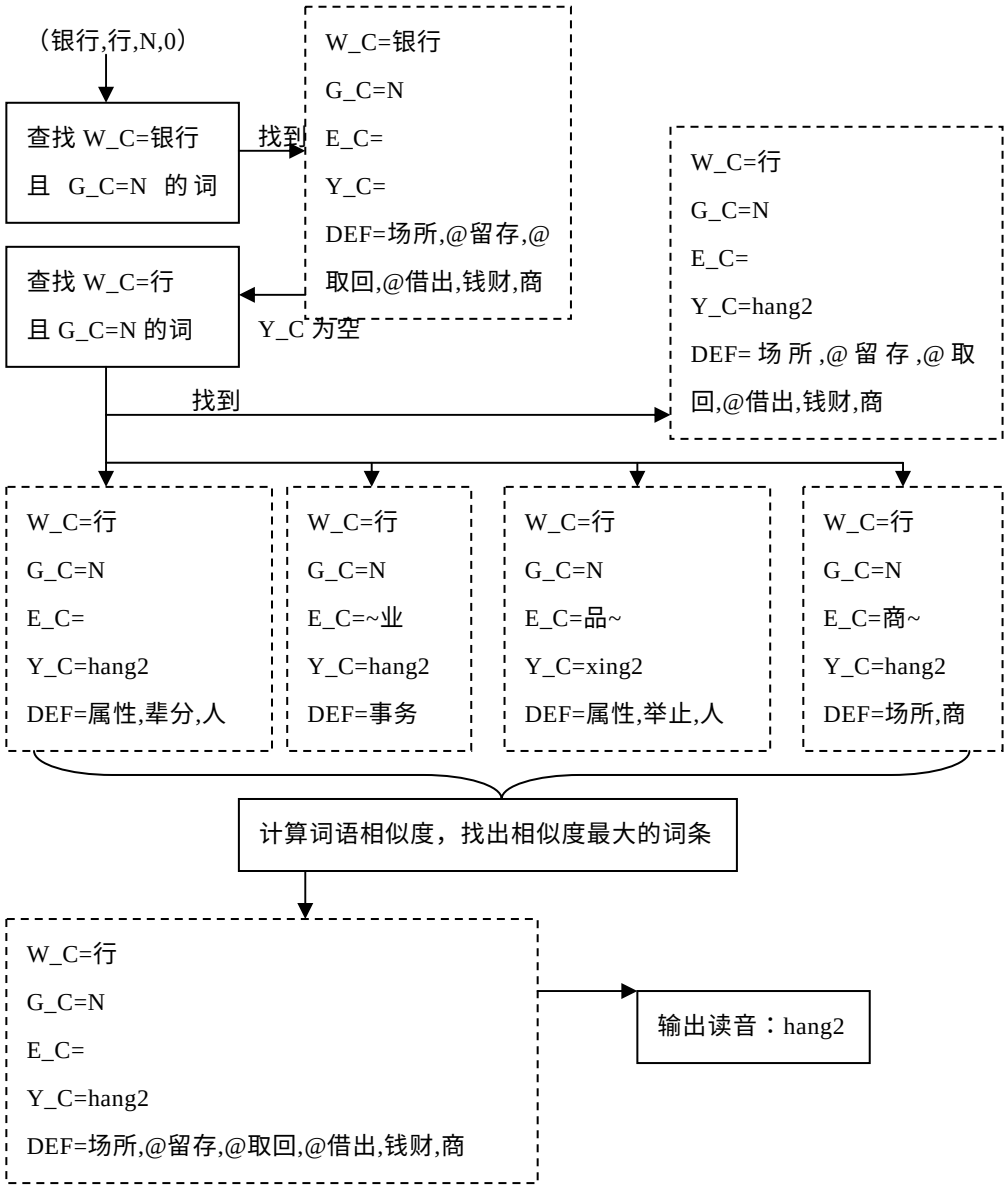


图 2-6 “ 银行 ”中“ 行 ”的读音判断过程

Figure 2-6 Syllable Determining Flow of “行” in “银行”

$$正确率=\frac{标注正确的词数}{多音字所在词总数}\times 100\%$$

从实验结果我们可以看出利用《知网》进行多音字处理的方法在单独运行时就能取得很好的效果，因此投入完整的字转拼音模块中可以使整体的正确率达到理想的实用效果。

这个算法有以下几个优点：

1. 脱离统计方法，在很大程度上避免了数据稀疏带来的问题；
2. 在解决新词和多音词读音判别方面有很强的鲁棒性；
3. 大大地减少了编写规则需要的人力、物力投入；
4. 方法简便、易于使用；

但它也有缺点：

1. 载入资源量过大，使算法运行效率不高；
2. 无法对虚词多音字进行处理（可以通过统计方法解决虚词多音字读音判断）。

2.5 本章小结

本章通过分析以往的多音字处理方法的优点与不足，提出运用语义计算来进行多音字处理。在详细介绍了《知网》这个语义资源之后，描述了利用《知网》通过计算语义相似度获取多音字读音的算法，并举例说明了算法的处理流程。最后应用较大规模的实验总结了这种方法的优缺点，证明了它的可行性，阐述了在文语转换系统的具体使用方法。

第3章 韵律短语边界识别

3.1 韵律短语

汉语由音节组词（短语）、由词构成语句。朗读或说话时，发音人将按词、短语插入长短不同的停顿，也会根据句子的结构和对句子的理解，不断变化语速、轻重、高低。韵律是超音段特征，表现为语音的时长、音高、音量的改变[20]。停顿是韵律特征之一，在连续语流中停顿的插入跟语句长、语义、语气、重音、语速等因素有关。于是一个汉语句子就被分成了若干组块，也就是韵律短语，所以通常将韵律短语定义为设定的两个停顿之间的短语和词[21]。韵律短语的边界就是停顿插入处。韵律短语边界识别的正确性将直接影响文语转换的自然度。

3.2 韵律短语边界识别

文语转换系统必须能够像人一样自动地决定在句子中何处插入停顿，因而韵律短语边界的自动识别必须拥有很高的效率和正确性。很多文语转换研究者都是把韵律短语边界识别与其它韵律参数一起进行预测，采用的方法有神经网络、隐马尔可夫模型、分类树等等，国外在这方面的研究取得了一些很好的结果[22]。有两位学者在一篇文章中论述了采用五种不同的方法从文本入手进行韵律短语边界识别情况，并通过实验对这五种方法的效果进行了比较。这五种方法都是基于分词词性标注序列的，分别是[23]：

1. trigram probabilities：对所有数据集中的三元序列进行统计，计算它们后面、前面、中间加入停顿的概率。测试时考察句子中每一个三元序列的概率是否大于某个阈值。如果大于这个阈值则加入停顿标记；
2. trigram probabilities + distance：在第一种方法的基础上引入距离，即当前位置离前一个韵律短语边界越远，它是下一个韵律短语边界的可能性越大；
3. trigram probabilities + distance + lookahead L trigrams：从整个句子中的全部三元文法概率考虑，不仅对当前三元序列选择概率最大的短

语标记，同时还观察后续的 L 个三元序列，以保证标记不互斥；

4. 组合概率法：对句子分词序列的所有短语组合情况进行停顿标记概率统计，选择一种平均概率最大的短语组合形式进行标记；
5. 标点符号法：先以标点为界将句子划分成若干块，然后再运用第一种方法标注韵律短语边界。

这五种方法的开放测试比较结果为：

表 3-1 5 种方法开放测试结果

Table 3-1 Open Test Results of Five Methods

方法	韵律短语边界正确率
1	70%
2	68.5%
3	70%
4	72%
5	69%

封闭测试结果均能达到 90%左右，可以看出统计的方法已经能够取得很好的结果，但计算量非常大。

国内一些研究者也模仿国外的算法进行了汉语文语转换韵律预测研究，但结果并不理想，且更没有单独对韵律短语边界识别的评价。

在自然语言理解中，语法短语是根据词法、句法来划分的。但韵律短语不一定与汉语语法短语一致。韵律短语可能是一个名词（体词）短语、动词（谓词）短语、介词短语等等，也可能是由一个语法短语和其前导成分或后继成分所组成。在一个韵律短语内部一般具有相对固定的韵律模式。在TTS的语句发音中，一个语句可以包括多个韵律短语，通常一个韵律短语在语句发音中连续读出，多个韵律短语之间往往可以插入长度适当的停顿，以增强节奏感。因此，韵律短语边界的识别也属于一种浅层句法分析过程。在浅层句法分析方法中有一种基于转换的规则学习方法，可以通过自学习获取规则用来进行句法分析。1995 年Eric Brill提出了一种叫做基于转换的错误驱动算法[24]，首先将其用于词性标注，得到的结果可以和统计的方法相媲美。而后有人把这种自学习方法用于识别英语中的基本名词短语，也同样取得了很好的效果[25]，并且方法简单、易用。国内有人将这种方法应用在汉语语法短语自动界定上[26]。因此，本文将采用此算法获取韵律短语边界识别规则。

3.2.1 基于转换的错误驱动算法介绍

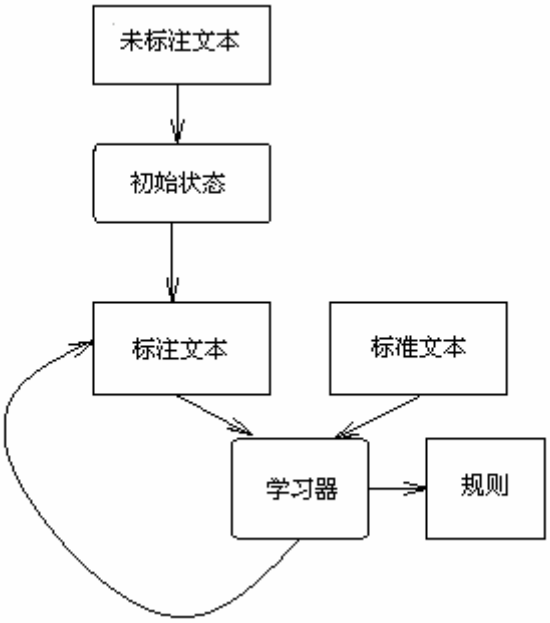


图 3-1 基于转换的错误驱动算法

Figure 3-1 Transformation-Based Error-Driven Learning

如图 3-1 所示，基于转换的学习方法以下列三部分资源为基础[27][28]：

1) 带标注的训练语料库；
2) 规则模板集合。规则模板集合用于确定可能的转换规则空间。
3) 一个初始标注程序。

基于转换的错误驱动学习算法是：

1. 初始标注。把训练语料中所有的标记去掉，用一个简单的初始标注程序标注出训练集中可能的标记。把这个结果作为系统的底线；
2. 生成候选规则集。在每个初始标注错误的地方，规则模板便用来生成候选规则，规则的条件就是词的上下文环境，动作就是改正错误标记所要做的动作；
3. 获取规则。把候选规则集中的每条规则分别运用于初始标注的结果，选出得分最高的规则（得分为正确的修改数减去错误的修改数得到的结果）。把这条规则运用于初始标注的结果作为下一轮循环的基础，并把这条规则作为规则序列中的第一条规则输出。重复以下过程直到得分最高的规则的得分为 0 或低于某个阈值为止：获取

候选规则集，给其中每条规则打分，选择得分最高的规则输出到规则集中，并把这条规则作用于当前语料库。

通过以上自学习过程就可以得到一个有序的规则集。

3.2.2 模板参数选择

汉语句子是由词语组成的。人们在朗读时通常会根据词语和句子结构插入停顿。那么在进行短语识别之前要先对句子进行分词、词性标注。目前，汉语文语转换系统中，汉语分词的基础是语法词典，短语合成也是在此基础上进行的。在自然语言理解中，语法短语是根据词法、句法来划分的。但韵律短语不一定与汉语语法短语一致。韵律短语可能是一个名词（体词）短语、动词（谓词）短语、介词短语等等，也可能是由一个语法短语和其前导成分或后继成分所组成。在一个韵律短语内部一般具有相对固定的韵律模型。

既然词性标注能够体现一定的句法结构，当然也可以用于识别韵律短语，因而把词性作为规则模板的第一个特征。英国爱丁堡大学的语音技术研究所的两位学者利用词性信息进行短语边界识别，封闭测试的正确率达到了90%以上[29][30]。另外，有实验说明，两停顿间的平均音节数为 3.6 个，一般认为一个呼吸群在七个音节左右[31]。可见，韵律短语边界与音节数量也有关系[32]，因此以音节数作为模板的第二个特征。

3.2.3 初始标注

在 3.2.1 中提到对训练语料进行初始标注的方法是去掉所有标记，然后用一个简单的标注程序去做初始标注。这里我们不再另做一个程序去进行初始标注，而是设定一个初始规则集，利用错误驱动的自学习程序根据初始规则集进行初标。这样就需要先获取初始规则集。初始规则集选择的好坏将直接影响自学习的收敛速度，因此本文采用简单统计的方法获得初始标注集。

已标注好的训练语料里所有韵律短语中出现频率最高一个或几个的词性和音节数用于组成初始规则集。

3.2.4 规则模板设计

规则模板，即规则的形式，在进行自学习算法之前必须预先指定，这样

自学习程序才能生成正确形式的规则。与其他学习算法不同的是，基于转换的错误驱动算法的模板非常简单，并不需要很多、很深的语言学知识，模板的数量也不需要很多。但是如果模板选择的不合适将严重影响自学习得到规则的可用性。

既然已经确定使用词性和音节数作为模板参数，下面需要做的就是决定模板的具体样式和排列形式。

尽管韵律短语边界不完全符合句子的语法结构，但人们在朗读时首先想到的是句子的逻辑结构，而这种逻辑往往又表现在语法结构上尤其是词性[33]。所以第一个模板为：

if 0:POS=X->PAUSE=*

“ 0 ”表示当前词，“ POS=X ”表示这个词的词性为 X，“ PAUSE=* ”表示这个词是否为韵律短语的结束边界词。当“ * ”被“ 2 ”替换时说明这个词是韵律短语结尾，当“ * ”被“ 1 ”替换时说明这个词不是韵律短语边界，当“ * ”被“ 0 ”替换时说明这个词后面不作停顿直接与后面词连读。

既然音节数是第二个模板参数，那么第二个模板应为：

if 0:POS=X&0:LENGTH=Y->PAUSE=*

其中“ LENGTH=Y ”表示当前词包含的音节数为 Y。当然不能仅仅考虑当前词而忽略上下文，因此在下面的模板中加入前一个词和后一个词的信息，就好像三元文法。于是得到所有模板：

(第1类) if 0:POS=X->PAUSE=*

(第2类) if 0:POS=X&0:LENGTH=Y->PAUSE=*

(第3类) if -1:POS=X&0:POS=Y->PAUSE=*

if 0:POS=X&1:POS=Y->PAUSE=*

(第4类) if 0:POS=X&-1:POS=Y&-1:LENGTH=Z->PAUSE=*

if 0:POS=X&1:POS=Y&1:LENGTH=Z->PAUSE=*

if 0:POS=X&0:LENGTH=Y&-1:POS=Z->PAUSE=*

if 0:POS=X&0:LENGTH=Y&1:POS=Z->PAUSE=*

(第5类)if 0:LENGTH=X&0:POS=Y&-1:POS=Z&-1:LENGTH=U->PAUSE=*

if 0:LENGTH=X&0:POS=Y&1:POS=Z&1:LENGTH=U->PAUSE=*

(第6类) if -1:POS=X&1:POS=Y&0:POS=Z->PAUSE=*

(第7类) if -1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U->PAUSE=*

if -1:POS=X&1:POS=Y&-1:LENGTH=Z&0:POS=U->PAUSE=*

if -1:POS=X&1:POS=Y&1:LENGTH=Z&0:POS=U->PAUSE=*

(第8类) if-1:POS=X&-1:LENGTH=Y&1:POS=Z&0:LENGTH=U&0:POS=V
->PAUSE=*
if -1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U&1:LENGTH=V->PAUSE=*
if -1:POS=X&1:POS=Y&-1:LENGTH=Z&0:POS=U&1:LENGTH=V
->PAUSE=*
(第9类) if-1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U&1:LENGTH=V&
-1:LENGTH=W->PAUSE=*

“ -1 ”表示前一个词，“ 1 ”表示后一个词。可以看出这些模板的覆盖范围是随着参数的增加而递减的。产生的规则也遵守这样的规律。鉴于这些模板的覆盖范围不同，将它们分成了 9 类，以便使用每一类模板都能够产生最少充分规则来纠正原有规则集标注时产生的错误。

3.2.5 韵律短语边界识别规则自学习

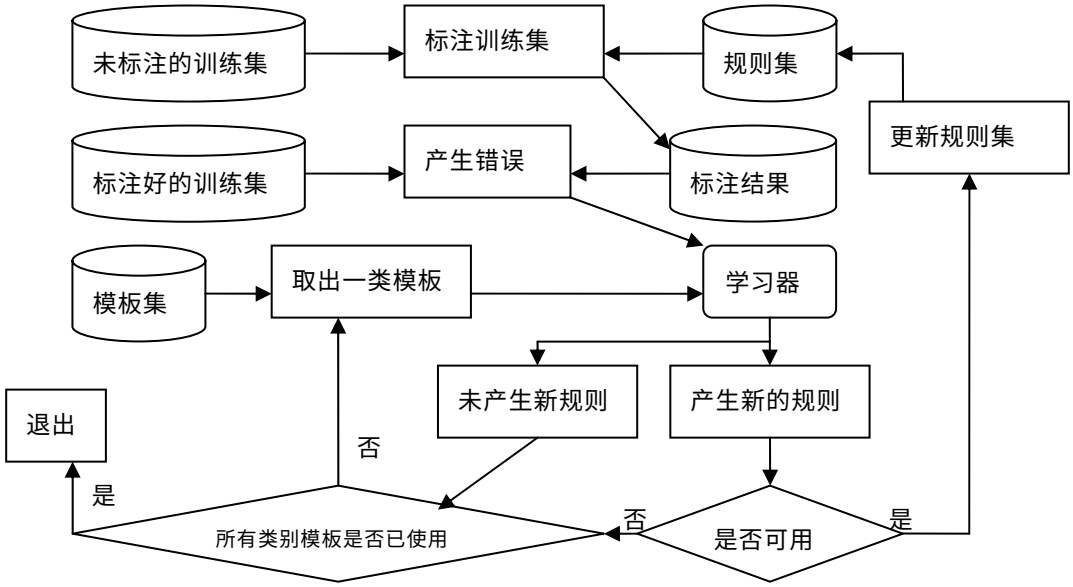


图 3-2 韵律短语边界识别规则自动获取流程

Figure 3-2 Automatic Acquisition Flow of Prosody Phrase Boundary Detection Rules

在得到了基于转换的错误驱动算法的资源即训练集以后，我们建立一个自学习系统来获取韵律短语边界识别规则（如图 3-2）。

一开始规则集里是预先统计出来的初始规则，运用这些初始规则去对训练文本进行韵律短语边界标注。将标注结果与标准训练集（即预先标注好的训练文本）进行比较，找出标注错误送入学习器根据模板进行规则获取。如果得到的规则所能纠正的错误数达到某个阈值并且不与原来的规则冲突，则将新规则添加到规则集中。然后再用更新后的规则集重新开始对训练文本进行标注。当得不到新的可用的规则时，从模板集中取下一类模板送入学习器用于生成新规则。这样往复进行直到所有的模板都已使用并得不到任何新规则时，算法输出规则集。

关于阈值的选取将在下一小节的实验结果与分析中进行详细说明。

3.3 实验结果与分析

3.3.1 实验语料建设

文语转换系统的目标是合成任何给定文本的语音进行输出，因此用来训练韵律生成模型的语料必须尽可能的覆盖所有的语言现象。我们建立了一个由 5725 个汉语普通话句子的文本和女声语音组成的语料库。这些句子包括简单、复合、陈述、疑问、感叹等句型。

为了给错误驱动算法提供标准训练集，人工根据观察波形、听辨语音对句子中韵律短语边界（即停顿）进行了标记。参与标记工作的都是经过专门训练的，对语音现象及声谱非常了解。下面举例说明一下标记情况，例句“五名死者包括一名妇女和两名儿童”。

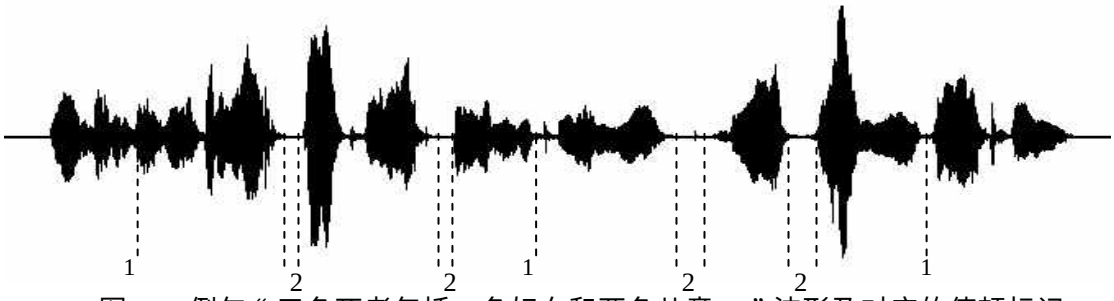


图 3-3 例句“五名死者包括一名妇女和两名儿童。”波形及对应的停顿标记

Figure 3-3 Waveform and Phrase Tagging of Example Sentence

从图中看到，标注“1”处是词语一级的停顿，也就是说该标记与前一个标记之间的一个或多个语法词汇朗读时都被看作一个词。如果其间包含多个语法词则这几个词之间的停顿标记为“0”。标注“2”表示韵律短语边

界。既然规则模板还是以语法词为主要基本单位的，那么在进行了上述的标记之后就需要对所有训练语句进行分词、词性标注。上面例句的分词、词性标注结果为“五/m 名/q 死者/nc 包括/vg 一/m 名/q 妇女/nc 和/c 两/m 名/q 儿童/ng”。词性标注符含义见附录表 A-1。

3.3.2 规则选取的阈值设定

上一节中对获得的规则是否可用做两种判断：一是判断新规则能纠正的错误数量是否达到阈值要求，二是判断新规则是否与规则集中的原有规则冲突。对于其中的阈值，不能仅凭主观臆断来定量，而是需要依靠实验来选取一个使韵律短语边界识别精确度较高又不影响整体性能的值。

把 5725 个句子的后 725 句作为开放测试集，用前 1000 句投入自学习系统来进行阈值选取实验。结果如下：

表 3-2 阈值选定实验结果

Table 3-2 Experiment Results of Threshold Choosing

阈值	精确度	得到的规则数量
1/3	91%	8858
9/24	91.5%	8357
5/12	92%	7274
11/24	91.3%	6524
1/2	91%	2091
2/3	87%	2091

阈值为新规则能够纠正的错误数和用规则集标注后的错误总数的比值。从表中可以看出，阈值选为 5/12 能够取得比较高的韵律短语边界自动识别精度。表中的精确度是封闭测试结果。

3.3.3 实验结果

根据上面选取的阈值，分别用前 1000 句和全部 5000 句训练文本投入自学习系统生成规则，然后进行封闭测试和 725 句开放测试，检验得到的规则对韵律短语边界识别的效果。

表 3-3 训练集不同的封闭及开放测试结果

Table 3-3 Close and Open Test Results of Different Training Data

训练集句子数	封闭测试精度	开放测试精度	得到的规则数量
1000	92%	73%	7274
5000	87.5%	77.1%	20400

3.4 TTS 中韵律短语边界识别规则的组织

从上一节的实验结果中可以看出虽然韵律短语边界识别的精确度较高，能够达到实用要求，但得到的规则数量是很庞大的。为此必须建立一套良好的规则组织机制，使匹配规则的效率能够达到实用要求。从规则模板的形式来看，不同的模板之间具有一种层次关系：从当前词开始增加参数并向前后词延伸。于是可以用“树”形结构来组织得到的规则。下面通过一个具体实例来说明规则的组织形式。

假设有如下几条规则：

if 0:POS=ut->PAUSE=0

if 0:POS=ut&0:LENGTH=2->PAUSE=0

if 0:POS=ut&0:LENGTH=2&1:POS=nd->PAUSE=2

if 0:LENGTH=2&0:POS=ut&1:POS=nd&1:LENGTH=4->PAUSE=2

if 0:POS=ut&1:POS=j->PAUSE=1

if 0:POS=ut&1:POS=vg->PAUSE=2

if 0:POS=ut&1:POS=nd->PAUSE=2

在 TTS 中的组织形式如图 3-4。

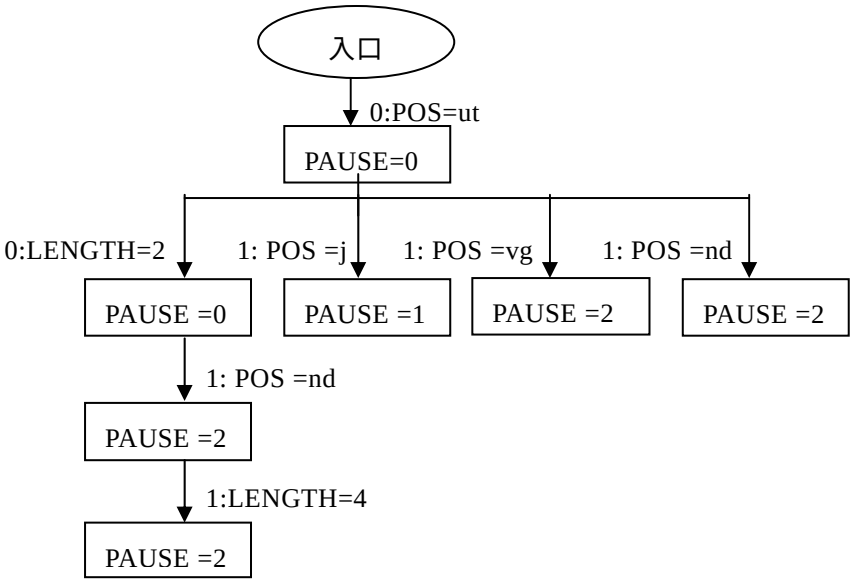


图 3-4 规则的树形组织举例

Figure 3-4 Example of Tree-like Organizing of Rules

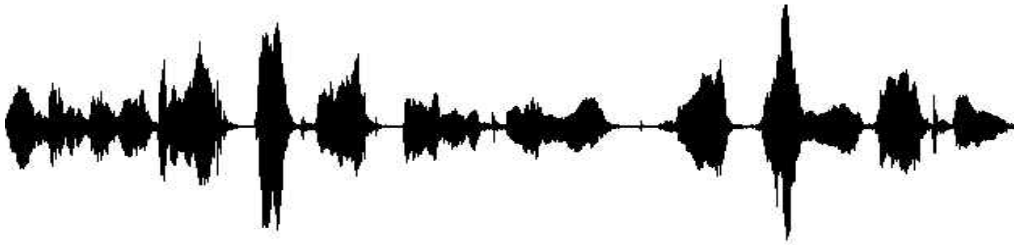
这株树与决策树有点类似，但它的树节点上没有对应的概率值。自学习

器得到的所有规则都可以用这种方式在 TTS 系统中形成一株树。在 TTS 系统中查找一条最合适的规则用来识别韵律短语边界时，对规则树进行先深搜索。这样尽管规则的数量很庞大，但树形组织大大削减了规则匹配时间，提高了系统运行效率。

3.5 本章小结

通过介绍韵律短语及其在汉语文语转换系统中的作用，引入本章的主要目标——自动识别韵律短语边界。借鉴以往在句法分析方面运用的基于转换的错误驱动算法，经过改造将其用来进行韵律短语边界识别规则的自学习上。另外，又提出使用树形组织文语转换系统中的韵律短语识别规则，使系统达到了很好的实用结果。

例如，本章中提到的句子“五名死者包括一名妇女和两名儿童”，经过韵律短语识别得到 5 个短语。它们是“五名死者”、“包括”、“一名妇女”、“和”、“两名儿童”。系统合成输出该句的波形和由朗读者直接朗读此句的波形如图 3-5。



（图 3-5-a）朗读者朗读例句的波形

（图 3-5-b）TTS 合成输出例句的波形

图 3-5 朗读与合成的比较

Figure 3-5 Comparison between Real Speech and Synthesized Speech

通过比较可以看出，本章所述的韵律短语边界自动识别方法在系统中取

得了很好的效果。

第4章 朗读重音判别

4.1 朗读重音

在连续语流中，各音节的响亮程度并不完全相同，有的音节听起来比其他音节重，简单地说，这就是重音。以词为考查对象，音位学可划分为正常重音、对比重音和弱重音。人们在口语交流中，常把在表情传意方面较重要的词读得重些，把其余的词读得轻些。语句重音是指由于句子语法结构、逻辑语义或心理情感表达的需要而产生的句子中的重读音，它不同于词重音，因为词重音只出现在词结构中。语句重音一般分为三种:语法重音、逻辑重音、心理重音[34]。其中，语法重音是指在不表示什么特殊的思想和感情的情况下根据语法结构的特点而把句子的某些部分重读的现象。逻辑重音是为了表示特殊的思想和感情而把句子的某些地方读得特别重的现象，又可称为“强调重音”或“感情重音”。

语言学家通过一些研究指出了语法重音的主要规律[35]：

1. 一般短句子里谓语部分常读语法重音；
2. 名词前面的定语常读语法重音；
3. 动词或形容词前面的状语常读语法重音；
4. 动词后面由形容词或其他动词充当的补语常读语法重音；
5. 有些代词常读语法重音。

逻辑重音的规律很难把握，它与语义、感情的关系很难让计算机理解。比如“今天是星期天”，如果重音在“星期天”，则讲话者强调的是今天是星期天不是别的日子；如果重音放在“今天”，则讲话者强调的是今天而不是昨天是星期天。

上面所述的均是句子重音（accent）。而重音有不同层次不同类型，通常把韵律词（朗读时呼吸、停顿的最小单元）重音称为stress，把比它层次高的重音称为accent，其中特别地把句子重音称为nuclear accent（核心重音）。stress和accent统称为prominence[36]。由于本文作者对重音的研究刚刚开展以及本文篇幅有限，不可能对句子重音和韵律词重音一一作以分析，因此本章先对二字韵律词重音的规律作以分析，再运用第三章的基于转换的错误驱动算法获取韵律短语重音判别规则。

4.2 韵律词重音

并不是所有的语言都有韵律词重音（如日语、法语就没有）。韵律词重音的作用单位是音节。Agaach M.C. Sluijter等认为，从发音的角度来看，重读音节发音时用力较大，因此用力是韵律词重音发音生理方面的相关物[37]。但是，重读音节的这个单一的发音特征（用力）却对应着较多的声学相关物，韵律词重音声学参数主要有：时长（duration）、频谱倾斜（spectaltilt）、元音品质（vowel quality）和强度（intensity）。另一方面，从知觉的角度来看，上述声学参数对韵律词重音知觉的贡献是不一样的，时长的贡献最大，其次是频谱倾斜，然后是元音品质，最后是强度，研究者们认为，之所以强度对韵律词重音知觉的作用很小，是因为它很容易受到环境噪音的影响，而其它声学参数对环境噪音却有较强的抵御力。因此，听者对韵律词重音和句子重音带来的频率增加的性质是不一样的，因为它们对应着完全不同的发音特征。从信息的观点来看，言语交际是一个对于说者是编码、对于听者是解码的过程，韵律特征只是这种编码解码的一种方式。而言语交际能够有效地实现的一个必要条件是这个解码刚好是编码的逆过程，这在韵律词重音的产生和知觉中是有所表现的。韵律词重音对应的发音特征是单一的，即用力，但是它的声学表现却是多样的，这是一个从一到多的变化。而从知觉的角度来看，听者能够把这些声学参数有效地综合起来，得到一个单一的知觉结果，即重读。

韵律词重音的一个重要作用是赋予言语以节奏感。正是这个韵律词重音把言语从音节水平组织到高一层次的节奏单元水平。由于单词通常都是由一个或者一个以上的节奏单元组成，因此，单词的边界一般同时也是节奏单元的边界。所以，韵律词重音是听者进行单词切分的一个重要依据。Cutler和Butterfield[38]通过研究发现，以英语为母语的听者倾向于把重读音节看作单词的开头，而这种倾向在其它的关于单词边界声学相关物不明显时，对于听者是十分有用的。节奏是指强弱相间，来自发展心理学的研究表明，婴儿能够分辨节奏感强和节奏感差的听觉刺激材料，并对前者有较好的倾向性，而这对儿童的语言习得是非常有用的。研究者们还发现，甚至哺乳动物对节奏也表现出一定的敏感性。在一些情况下，当不能通过韵律词重音赋予言语节奏感时，往往会有一种特殊的accent代替韵律词重音赋予言语以节奏感。可见，节奏是言语的一个很强烈的要求。也正因为如此，节奏感成为韵律构词

学的一个重要依据。

4.3 分析和实验用语料建设

为了进行朗读重音的定性定量分析，语料是必需的。由于已有的语料库没有标记重音，为此需要重新加工语料——在原有语料的基础上标注重音。重音的标注必须通过对语料中实际语音的听辨来判断，因此是一个需要了解文语转换系统需求和语音学知识的人来参予的工作，还需要有一个便于使用的标注工具。下面分别介绍一下标注工具的制作和重音的标注。

4.3.1 标注工具的制作

在第三章的工作中已经对韵律词和韵律短语进行了标注，另外原语料的基本标注还有每个音节在句中的起始位置，因此可以通过将每个句子按照韵律短语、韵律词、字逐层拆分显示，并与音库中句子相应的语音段对应，这样不仅可以听全句，还可以听任何一个韵律短语、韵律词及字。标注工具的制作由一个本科生来完成，本文作者提供了相应的语音分段提取和拼接程序模块。标注工具用户界面参见附录图 A-1。

4.3.2 重音的标注

使用重音标注工具分层次对韵律短语、韵律词、字进行了重音标注。标注工具可以将标注结果以两种形式输出：

1) 句中嵌套式：在标好韵律短语边界的句子文本中直接添加重音标记。例如，“ 道琼斯股票指数创纪录。”

   这句话加标韵律短语后是“ 道琼斯#2 股票#1 指数#2 创纪录。”

   添加重音标记后是“ 道 S0 琼斯#2 股 S0 票#1 指 S0 数 S1#2 创 S0 纪录 S2。”

   “ S0 ”表示前面的字重读，即“ 道琼斯 ”的“ 道 ”、“ 股票 ”的“ 股 ”、“ 指数 ”的“ 指 ”、“ 创纪录 ”的“ 创 ”。

   “ S1 ”表示其前的韵律词重读，即“ 指数 ”。

   “ S2 ”表示其前的韵律短语重读，即“ 创纪录 ”。

2) 字列法：以字为单位，用 5 列数字分别标明其在韵律短语中的位置、韵律词中的位置、所在韵律短语重读情况、所在韵律词重读情

况、本字重读情况。仍是上面那个例子，重音标注输出形式：

道 1 0 0 0 1
琼 2 1 0 0 0
斯 3 2 0 0 0
股 1 1 0 0 1
票 2 2 0 0 0
指 3 1 0 1 1
数 4 2 0 1 0
创 1 0 1 0 1
纪 2 1 1 0 0
录 3 2 1 0 0

第一列为句中的每一个字，第二列的数字表示该字是所在韵律短语中的第几个字，第三列的数字表示该字是所在韵律词中的第几个字，第四列表示该字所在韵律短语是否重读，第五列表示该字所在韵律词是否重读，第六列表示该字是否重读，“ 0 ”表示不重读，“ 1 ”表示重读。

4.4 二字韵律词重音分析

二字词即双音节词中两个字的重读模式与很多因素有关，如词性、词义、构词方式等等。下面将根据对实验语料的简单统计数据对普通话双音节常用词重音音节的规律进行分析。实验语料的 1000 句文本经过分词得出双音节词为 2816 个。

4.4.1 分析

4.4.1.1 按构词方式分析[39]

1. 单纯词以第一个字重音为基本模式，例外的有：叠字单纯词中的副词，两个字均为平声时第二个字重读；
2. 带后缀和类似后缀的词素的合成词，75%第一个字重读，例如：“ 子 ”字作后缀的词有 41 个，全部是其前面的字念重音；“ 头 ”字作后缀的词有 21 个，其中 9 个念重音，其余是它前面的字重读。
3. 带前缀和类似前缀的词素的合成词，多数第二个字重读；
4. 动宾式合成词，92%第二个字重读；
5. 主谓式合成词，80%第二个字重读；

6. 补充式合成词，基本模式为第二个字重读；
7. 偏正式合成词，73%第二个字重读；

4.4.1.2 按词性分析

同一个双音节词，词性不同，前后字的重音情况不同。如下表所示：

表 4-1 词性分析重音规律

Table 4-1 Stress Rules Based on Part of Speech

二字词语	词性	重读音节位置	词性	重读音节位置
报告	动词	1	名词	2
出口	动词	1	名词	2
东西	方位词	2	名词	1
访问	动词	1	名词	2
多少	代词	1	形容词	2
记录	动词	1	名词	2
练习	动词	1	名词	2

实验语料中有 43 个词同时具有动词、名词两种词性，其中 30 个满足类似表中的规律：动词的第一个音节重读，名词的第二个音节重读。

4.5 韵律短语重音判别规则获取

第三章中成功地运用基于转换的错误驱动算法进行了韵律短语边界识别规则的自动获取，并在实际系统运用中取得了很好的效果。因此本节采用同样的方法进行韵律短语重音判别规则获取。

Alan W Black[40]在英语重音判别中使用词性作为算法参数，因此本文也采用词性作为规则模板的唯一参数，然后考虑上下文，即前一个词和后一个词的词性，因此模板集为：

if 0:POS=X->STRESS=*

if -1:POS=X&0:POS=Y-> STRESS =*

if 0:POS=X&1:POS=Y-> STRESS =*

if -1:POS=X&1:POS=Y&0:POS=Z-> STRESS =*

除“ STRESS ”表示是否重读之外，其它标记同第三章。“ STRESS＝0 ”表示不重读，“ STRESS＝1 ”表示重读。

初始规则集的形成以及自学习算法均同第三章，这样将实验语料中的前 900 句作为训练语料，后 100 句作为测试语料。判别规则是否有用的阈值仍取 5/12，则得到 3025 条规则。封闭测试正确率为 46%，开放测试正确率结

果为 19%。正确率很低，无法在 TTS 系统中实用化。

4.6 本章小结

本章介绍了朗读重音的概念和分类，以及语法重音的规律。然后引入韵律词重音，对其产生和感知机理进行了详细介绍，并通过简单统计对二字韵律词的重读音节规律进行了分析。最后利用第三章韵律短语边界识别规则的自学习机制进行了韵律短语重音判别规则的获取，用以自动判别朗读时韵律短语是否重读，其实验结果并不理想，不能用于实际系统。

结论

近年来，语音对话系统、呼叫中心、语音触发的网站和电子邮件服务等实际应用的迅速发展，掀起了对文语转换技术的需求的一个前所未有的高峰。大量的应用需求也促使 TTS 技术的研究和开发迈上了一个新的台阶。虽然合成语音的质量已达到了基本可以接受的水平，其自然度与人的语音还有相当大的差距。其中在节奏、轻重、停顿等方面处理不当，使合成语音听起来很别扭。本文从字音转换、韵律短语边界识别、重音判别入手，对汉语语音合成自然度改善进行了研究。

《知网》作为目前国内汉语较为完备的语义资源为本文利用语义计算进行多音字处理提供了一个可计算、可推演环境。经过查穷举表、匹配词性规则、语义相似度计算三个步骤，TTS 字音转换正确率明显提高，并解决了新词和多音词读音的判断难点。

基于转换的错误驱动算法是一个经典的句法分析算法。本文运用该算法建立了一个韵律短语边界识别规则自学习系统，从语料中获取了大量规则，经过树形结构组织在 TTS 实用效果斐然。

汉语是声调语言，是否存在固定的词重音模式还没有定论。重音有多少种类、有多少级别、如何判定重音也没有定论。直接从文本中推断重音音节的位置也是一个难点。本文对二字韵律词的重音规律进行了分析，得出了一些粗糙的数据。而利用自学习获得的韵律短语重音判别规则的实用性很差。尽管如此，本文作者始终认为，汉语重音的研究非常重要，用好了重音信息一定会对提高合成语音自然度大有帮助。因此，需要在今后的研究中更加深入。

总之，本文在改善汉语语音合成自然度方面的研究取得了一些可喜的成果，并有技术可以直接应用到实际系统中。但所关注的毕竟仅是文语转换中很小的一部分，要使合成的语音能够像人的一样自然还需要更深更广的研究。

参考文献

1 王仁华．语音合成技术及国内外发展现状．中国呼叫中心产业发展研究报告.2000

2 Min Chu，Hu Peng. A Objective Measure for Estimating MOS of Synthesized Speech. Best paper award from COCOSDA，Eurospeech2001， Aalborg， Denmark， 2001

3 A.Batliner，R.Kompe，A.KieBling，M.Mast，H.Niremann，E.Noth. M=Syntax+Prosody：A Syntactic-Prosodic Labeling Scheme for Large Spontaneous Speech Databases. Speech Communication. 1998(25)

4 初敏．韵律研究与合成语音的自然度．第五届现代语音学学术会议文集，2001

5 John Goldsmith. Dealing with Prosody in a Text-to-Speech System. International Journal of Speech Technology. 1999(3)

6 蔡莲红，吴志勇，王玮，陶建华，王志明．语音技术的拓展与展望．计算机世界．2001

7 吴志勇，蔡莲红．语音合成技术的原理．中国呼叫中心产业发展研究报告．2000

8 Min Chu，Hu Peng，Eric Chang. A Concatenative Mandarin TTS System Without Prosody Model and Prosody Modification. Proc. of 4th ISCA Tutorial and Research Workshop on Speech Synthesis，Perthshire，Scotland，2001

9 Min Chu，Hu Peng，Hongyun Yang，Eric Chang. Selecting Non-Uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer. ICASSP2001，Salt Lake City，2001

10 蔡莲红．TTS系统的研究与发展．http://www.ctiforum.com/

11 荀恩东，蔡萌，李生，赵铁军．基于时域基音同步叠加方法文语转换系统．王承发．第五届全国人机语音通讯学术会议，哈尔滨工业大学，1998：哈尔滨工业大学出版社，1998：318－322

12 潘以锋．计算机在汉字自动注音中的应用．上海师范大学学报（自然科学版）．1996，第 25 卷（第 4 期）

13 孙玉琦，张凯，王晓龙，徐志明．基于规则和统计相结合的多音字研究．

王承发．第五届全国人机语音通讯学术会议，哈尔滨，1998：哈尔滨工业大学出版社，1998：323－326
14 周强，冯松岩．构建知网关系的网状表示．2000 年多语种信息处理国际研讨会，乌鲁木齐，2000：新疆大学出版社，2000：139－145
15 刘晓军．汉语语义资源建设的研究与实现．哈尔滨工业大学硕士论文．2001：1-2
16 董振东，董强．知网． http://www.keenage.com/
17 LV Xin, LIU Zhan-yi, ZHAO Tie-jun, WANG Li. Dealing with Polyphone in Text-to-Speech System Using How-Net. 徐明星．第六届全国人机语音通讯学术会议，深圳，2001：清华大学出版社，2001：159－162
18 杨晓峰，李堂秋，洪青阳．基于实例的汉语句法结构分析歧义消解．http://www.keenage.com/
19 赵铁军，吕雅娟，于浩，杨沐韵，刘芳．提高汉语自动分词精度的多步处理策略．中文信息学报．2001，第 15 卷（第 1 期）
20 Terken , Collier. The Generation of Prosodic Structure and Intonation in Speech Synthesis. Speech Coding And Synthesis. 1995
21 Sangho Lee and Yung-Hwan Oh. Tree-based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems. Speech Communication. 1999
22 Steven Abney. Prosodic Structure, Performance Structure And Phrase Structure. Speech and Natural Language Workshop , San Mateo, CA, 1992：Morgan Kaufmann Publishers, 1992：425－428
23 Eric Sanders, Paul Taylor. Using Statistical Models to Predict Phrase Boundaries For Speech Synthesis
24 Eric Brill. A Corpus-Based Approach to Language learning. Elsevier Science. 1993
25 Claire Cardie, David Pierce. Error-Driven Pruning of Treebank Grammars for Base Noun Phrase Identification. Proceedings of ACL-Coling. San Francisco, U.S.A., 1998
26 周强.一个汉语短语自动界定模型．软件学报．1996，第 7 卷增刊：315－322

27 孙宏林，俞士汶．浅层句法分析方法概述．当代语言学．2000，第 2 期
28 Eric Brill. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Elsevier Science, 1995
29 Alan W Black, Paul Taylor. Assigning Intonation Elements And Prosodic Phrasing For English Speech Synthesis From High Level Linguistic Input. http://citeseer.nj.nec.com/. 1996
30 Alan W Black, Paul Taylor. Assigning Phrase Breaks From Part-of-Speech Sequences. Eurospeech97, Rhodes, Greece, 1997：995－998
31 应宏，蔡莲红．基于助词驱动的韵律短语界定的研究．中文信息学报．1999
32 Yao Qian, Min Chu. Segmenting Unrestricted Chinese Text into Prosodic Words Instead of Lexical Words. ICASSP2001, Salt Lake City, USA, 2001
33 Xin Lv, Tie-jun Zhao, Zhan-yi Liu, Mu-yun Yang. Automatic Detection of Prosody Phrase Boundaries for Text-to-Speech System. Harry Bunt. 7th International Workshop on Parsing Technologies, Beijing, 2001：Tsinghua University Press, 2001：135－141
34 王玮，蔡莲红．数据挖掘走进语音处理．计算机世界．2001，6 月
35 胡裕树．现代汉语（增订本）．上海教育出版社，1981：147－159
36 仲晓波，杨玉芳．国外关于韵律特征和重音的一些研究．心理学报．1999，第 31 卷（第 4 期）
37 Agaath M C Sluijter, Vincent J van Heuven. Spectral Balance as an Acoustic Correlate of Linguistic Stress. J Acoust Soc Am. 1996, 1000 (4)：2471－2484
38 Catharine H Echols. The Perception of Rhythmis Units in Speech by Infants and Adults. Journal of Memory and Language. 1997, 36：202－225
39 殷作炎．关于普通话双音常用词轻重音的初步考察．中国语文．1982，第 3 期：168－171
40 Alan W Black. Comparison of Algorithms for Predicting Accent Placement in English Speech Synthesis. Proceedings of the

Acoustics Society of Japan. 1995, Vol.4 (1)：275－276

附录

TTS 中词性标注符与 How-Net 中词性标注符映射关系如下表：

表 A-1 TTS 与 How-Net 词性对照表

Table A-1 The Contrast Table of part of speech tag in TTS and How-Net

TTS	说明	How-Net	说明	TTS	说明	How-Net	说明
ng	一般名词	N	名词	vz	助动词	AUX	助动词
nx	中文姓氏	N	名词	vq	趋向动词	V	动词
nm	中文人名	N	名词	vb	补语助动词	V	动词
nd	地名	N	名词	a	形容词	ADJ	形容词
ny	外国译名	N	名词	z	状态词	ADJ	形容词
nj	机构名	N	名词	d	副词	ADV	副状
nq	行政单位名	N	名词	p	介词	PREP	介词
nc	称呼词	N	名词	used	助词“的”	STRU	助词
t	时间词	N	名词	usdi	助词“地”	STRU	助词
s	处所词	N	名词	usdf	助词“得”	STRU	助词
f	方位词	N	名词	ussu	助词“所”	STRU	助词
m	数词	NUM	数词	ut	时态助词	STRU	助词
q	量词	CLAS	单位词	o	象声词/拟声词	ECHO	拟声词
b	区别词	ADJ	形容词	e	叹词	ECHO	拟声词
r	代词	PRON	代词	h	前接成分	PREFIX	前缀
vg	一般动词	V	动词	k	后接成分	SUFFIX	后缀
vx	系动词	V	动词				

369 个常用多音字按音序排列如下：

表 A-2 常用多音字列表

Table A-2 Common Polyphone List

阿	乘	逢	划	勒	摩	丧	校	正
挨	澄	缝	还	乐	抹	扫	肖	只
艾	匙	佛	晃	累	那	色	芯	中
扒	冲	否	会	擂	难	刹	兴	种
把	仇	服	混	棱	囊	煞	行	重
耙	臭	脯	豁	俩	泥	扇	畜	爪
柏	处	嘎	几	凉	宁	上	旋	转
百	揣	盖	济	量	拧	少	血	椎
般	传	干	夹	撩	弄	舍	压	着

膀	创	杆	贾	燎	哦	省	哑	兹
磅	绰	岗	假	了	沤	盛	咽	钻
蚌	刺	镐	监	裂	胖	石	燕	作
剥	撮	胳	间	淋	刨	什	要	茜
薄	答	葛	见	馏	炮	识	掖	莞
堡	打	格	将	笼	泡	数	椅	拗
背	大	蛤	浆	隆	劈	说	荫	捋
奔	待	个	降	搂	片	伺	殷	叨
绷	担	给	嚼	露	漂	似	饮	呷
辟	单	更	角	陆	撇	宿	应	呱
臂	弹	供	教	率	屏	遂	佣	咧
扁	当	勾	桔	绿	迫	踏	与	唠
便	挡	估	结	论	铺	倘	吁	嗑
别	倒	骨	解	罗	仆	趟	约	嘀
瘪	得	谷	禁	落	朴	提	钥	嚎
并	的	冠	劲	络	瀑	挑	晕	沌
伯	地	观	颈	麻	奇	帖	载	孱
泊	调	龟	卷	吗	呛	通	攒	杈
卜	丁	柜	倔	埋	强	吐	脏	柚
参	钉	过	觉	脉	翘	驮	择	栎
藏	斗	汗	咖	蔓	切	哇	曾	槟
叉	都	好	卡	氓	茄	瓦	扎	腌
查	读	号	咯	猫	亲	乌	轧	臊
差	肚	喝	看	冒	区	洗	栅	禅
拆	度	荷	扛	没	曲	系	炸	沓
颤	顿	和	坷	闷	圈	厦	粘	衩
场	囤	横	壳	蒙	雀	吓	占	裨
长	恶	哄	空	靡	嚷	鲜	涨	蛄
朝	发	红	拉	秘	任	纤	召	簸
吵	坊	糊	喇	泌	撒	相	折	酊
车	菲	哗	姥	模	塞	巷	蛰	蹊
称	分	华	烙	磨	散	削	挣	啰

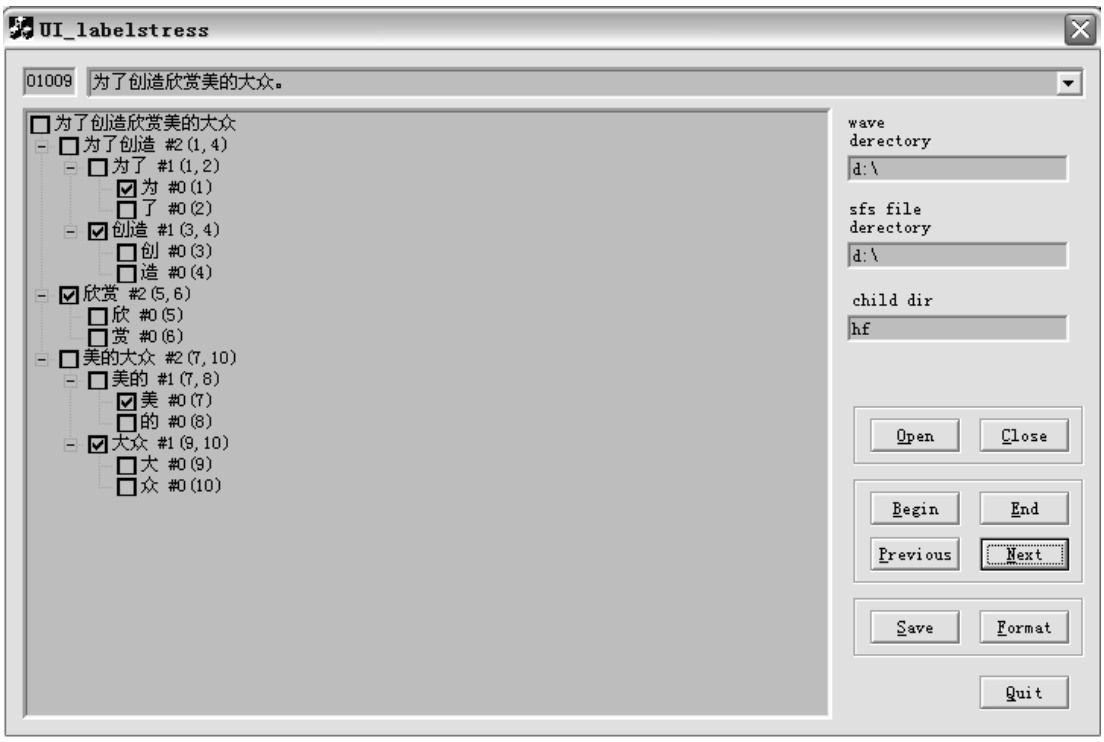


图 A-1 重音标注工具界面

Figure A-1 The Surface of Speech Stress Labeling

攻读学位期间发表的学术论文

[1] LV Xin, LIU Zhan-yi, ZHAO Tie-jun, WANG Li.Dealing with Polyphone in Text-to-Speech System Using How-Net.徐明星.第六届全国人机语音通讯学术会议，深圳，2001：清华大学出版社，2001：159－162

[2] Xin Lv，Tie-jun Zhao，Zhan-yi Liu，Mu-yun Yang.Automatic Detection of Prosody Phrase Boundaries for Text-to-Speech System.Harry Bunt.7th International Workshop on Parsing Technologies，Beijing，2001：Tsinghua University Press，2001：135－141

致谢

在攻读硕士学位期间，机器翻译实验室的老师和同学都曾给予我很大帮助，尤其是在语料的建设、词典的修正这些需要大量人力、物力的工作上。因而，首先要感谢实验室全体两年来对我研究的支持和帮助。

特别感谢的是我的指导老师赵铁军老师，是他引导我进入了一个崭新的研究领域，让我有更广阔的施展才智的空间。他对待工作的认真态度、为人的谦逊都是我学习的榜样。

其次要特别感谢的是刘占一师弟，他帮助我完成了很多实验，并给我的研究提供了很多崭新的思路。

感谢李生老师和于浩老师在论文撰写指导思想上的帮助，感谢杨沐昀老师在论文撰写方面给予我的帮助，感谢姚建民老师帮我寻找语料建设人员，感谢吕亚娟、孟遥两位师姐在算法上的指导。

感谢我的学友和朋友们对我的关心和帮助。

另外，还要感谢微软研究院的初敏博士。在微软中国研究院实习的三个月期间她教会了我很多知识，并对我回校之后提出的一些问题做了细致的解答。

总之，本论文工作期间得到了多方面的帮助，这里我再一次向他们表示最由衷的感谢！

1

3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33
34
35
36
37
38
39
40