汉语 TTS 中基于语音词的基元选取

吴志勇 蔡莲红

清华大学计算机科学与技术系 北京 100084

wuzy@tts.cs.tsinghua.edu.cn clh-dcs@tsinghua.edu.cn

摘要：在语音合成系统中，以往的分词处理，均以语法为基础。但以语法知识形成的节律与语音节律不尽相同。经分析研究，我们认为，在 TTS 系统的研究中，提出“语音词”（Speech word，简称 SpWord）的概念是必要的。本文将讨论语音词的定义；汉语 TTS 系统中基于语音词的基元选取；以及基元选取策略、搜索算法等。

一、引言

在语音合成系统中，语音的节律是非常重要的，它直接影响语音的可懂度和自然度。以往的分词处理，均以语法为基础。但以语法知识形成的节律与语音节律不尽相同。经分析研究，我们认为，在 TTS 系统的研究中，提出“语音词”（Speech word，简称 SpWord）的概念是必要的。语音词是指从语音学的角度来界定的词。

本文将讨论语音词的定义；汉语 TTS 系统中基于语音词的基元选取；以及基元选取策略、搜索算法等。

二、语音词

1 为什么要提出“语音词”的概念

通常一个人说话时，不可能一口气说上一长段话，也不是按音节一顿一顿地来说。实验标明，一段连续没有停顿或停顿不当的语流，其可懂度比具有正常轻重和节奏语流要下降很多。那么，如何确定比音节大，比语句小的语音单位呢？

语言学家认为这样的单位是：词或短语。但是对于什么是“词”，专家并没有给出便于操作的说一不二的结论[1]。我们的研究发现，按语法知识形成的节律与语音节律不尽相同。仅以语法来界定词或短语[2]来合成语音是不理想的。

胡明扬给出的调查结果表明：普通人与语言学家对词的认识不太相同，但相当一致的是“语感”。王洪君[3]认为构词法分为语法和语音两个方面。从语音的角度来看构词，汉语多采用加和构词法。通过对大量自然语流的研究，我们认为提出“语音词”（Speech word，简称 SpWord）的概念是必要的。人在说话时，以“语音词”为基本节奏单元，以韵律短语为呼吸单元。在“语音词”之内，各个音节之间的联系是比较紧凑，语流连续自然；而在不同的韵律短语之间，则由于较长的停顿而相对比较独立。

根据人说话时发音的这种特点。这种两次较短停顿之间的连读语音，就可以称之为“语音词”。在语音合成时，首先以“语音词”为单位进行音节基元的选取。再辅之以短语、句子的上下文相关信息，对“语音词”进行精选和拼接。按照这种方法，“语音词”听起来就会比较自然、流畅；整个语句也有较好的效果。

2 语音词在 TTS 中的地位

语音词是指从语音的角度来界定的词。语音数据库中，我们是根据人的语感来划分词的，称之为语音词。在 TTS 系统中，由文本分析模块提供分词结果。因此在这里我们不讨论语音

词的获得，仅介绍语音词在 TTS 系统中的地位。

语音词是连续语流中的一种切分单位，是语句节律的体现。从语音角度，它携带了音位、声调、轻重、长短等信息。从语法角度，它携带了词性、语义、语用等信息。从感知的角度，语音词是感知理解的基本单位。所以语音词对合成语音的可懂度和自然度至关重要。

冯胜利 [4] 探讨了汉语的韵律词（Prosodic Word），给出了“韵律级层”的概念。韵律构词学中的韵律系统分为四个“级层”：韵律词、音步（foot）、音节（syllable）、韵素。并认为汉语最基本的音步是两个音节 [4]。在我们的研究中，对自然语句进行切分和标注，从大到小的单元依次是：语句、块（chunk）、语音词、音节。合成时，我们从语音数据库中挑选音节，由相邻音节组合成语音词，进而构成块或语句。

三、TTS 系统中基元的选取

在基于拼接的 TTS 系统中，主要工作是在已知文本序列的条件下找出最适当的语音基元，然后拚接成连续语流。过去，许多的汉语 TTS 系统，其波形基元多为音节。而音节来自词语或负载句的特定位置，因此拼接出的语流没有语调信息，听起来呆板，平淡无味。在基于语音数据库的 TTS 中，语音基元来自数据库中的大量语句。这些语音基元隐含了音段特性和韵律特性，因此拼接出语流的自然度大为提高。但是在这样的系统中，基元选取是个很复杂的问题。

设与给定文本对应的音节序列为：$y_1, y_2, ...y_j ...y_n$，这里 j 为音节序号。

选择音节时，除确保读音正确外，还应考虑候选音节的声学参数和它所处的语境。因此，音节的特征集可表示为：

$C = \{L, A, P\}$，其中 L 表示语言学参数、A 表示语音学参数、P 表示位置参数。如语义、词性、基频、时长、幅度、音节在词中或句中的位置等。因此，基元的选取应考虑语言学和语音学规则，还要考虑改音节的声学参数。

通常，在语音数据库中，任一个音节都备有多个候选语音数据，如：$y_{j1}, y_{j2} ... y_{jk}, ... y_{jm}$，$y_{jk}$ 表示第 j 个语音基元的第 k 个候选者。

选择基元时，从 Bayes 的原则出发，计算每个候选基元特征的距离函数，选中其距离最小者，$F_{y_{jc}} = \min\{F_{y_{j1}}, F_{y_{j2}} ... F_{y_{jk}}, ... F_{y_{jm}}\}$。则 y_{jc} 为候选基元。

从语音库中，可以得知当前音节的音段参数、位置参数、与前后音节耦合度；相邻前后音节参数；计算得到它们的韵律参数；文本的韵律分析可以给出音节所在语音词的参数以及语句参数。基元选取的任务就是根据文本韵律分析对语音基元的要求和语音学的规则，选择“理想”的基元，拼接出连续自然的语流。

四、基于语音词的基元选取算法

1 基于语音词的基元选取算法的提出

我们为 TTS 系统设计了大量的短句，建立了语料库。对语料库中的语音信息进行了分析、标注 [5]。在合成时，充分利用语音库中原始信息，以使得合成结果的韵律、自然度等都能够取得比较好的效果。

当需要从语音数据库中选取适当的基元时，理想的情况是语音库中包含有“理想”的音

节，这时可采用“绝对匹配”的方法找到它。

但问题是如何确定“绝对匹配”的长度？如果采用语句级的匹配算法，其搜索时间太长。实验结果表明，当音节库含有一万五千个音节，语句长度平均为 10 个汉字时，其平均搜索时间 大约为 2 分钟。对于语音合成系统的实时性要求来说，这显然是不可容忍的。因此必须探求更好的语音基元搜索算法。其实，语料库中也不可能存在完全匹配的音节。

另一方面，在连续语流中，词内相邻音节的协同发音较为严重，词间次之；因此可以重点考虑语音词内部基元的选取。

另外，人们在听辨时，要求语音词内部相对比较紧凑和连续，语音的自然度高、轻重对比明显；而语音词之间，则要求其间有一定的停顿，连续性要求相对比较低一些。

针对上述情况，将选取基元的匹配策略确定为：减小 n-gram 模型的维数，基元精选的算法限制在一个语音词内。这样缩短了匹配的长度，计算的复杂度和计算量大为减小。合成语句时，再增加一些辅助的考虑，基本能满足系统实时性和语音质量的要求。我们称之为基于语音词的滑动窗口匹配算法。

2 基于语音词的基元选取算法描述

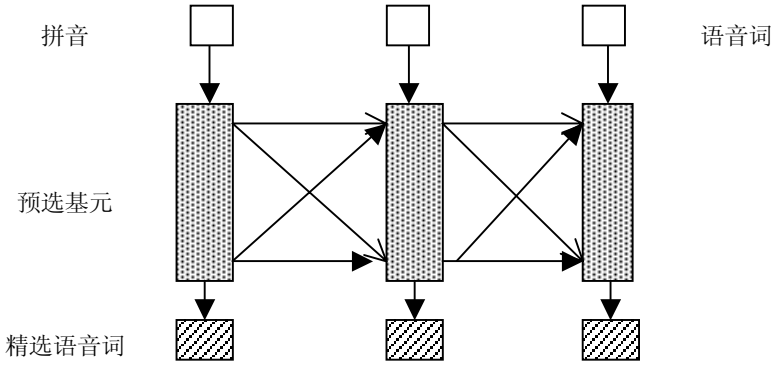

图 1：基于语音词的基元选取算法

选取算法包括两个部分，语音词内部的基元预选，以及语音词候选基元的精选

基于语音词的基元选取算法主要由两步完成：一是语音词内部的基元的预选，二是语音词的精选（即从不同基元搭配成的语音词中选出最优者）。我们首先根据文本分析的结果，考虑语音词内部的基元选取，在确定了语音词内部的所有候选单元以后，再根据短语、句子等的上下文信息考虑各候选基元的进一步精选，最终确定各个语音词的语音基元，从而确定整个合成单元的基元（图 1）。

(1) 语音词内部的基元选取

基元选取算法首先是预选语音词内部的基元。

在上文中已经介绍了基元选取所涉及和必须考虑的一系列语音及声学参数，而这些参数对于语音词内部基元的选取以及语音词的精选其作用并不是完全等同的。可以根据这些参数和语音词的关系，而分成两类，分别进行考虑。

在语音词内部基元选取的过程中，由于我们将基元的选取限制在一个语音词的内部，所以可以只考虑和语音词内部相关联的语音和声学参数，主要有：音节耦合度，前后音声调模型，前后音声韵信息，音节在词中位置，音节的韵律参数等。

语音词内部基元选取的基本流程如图 2 所示。

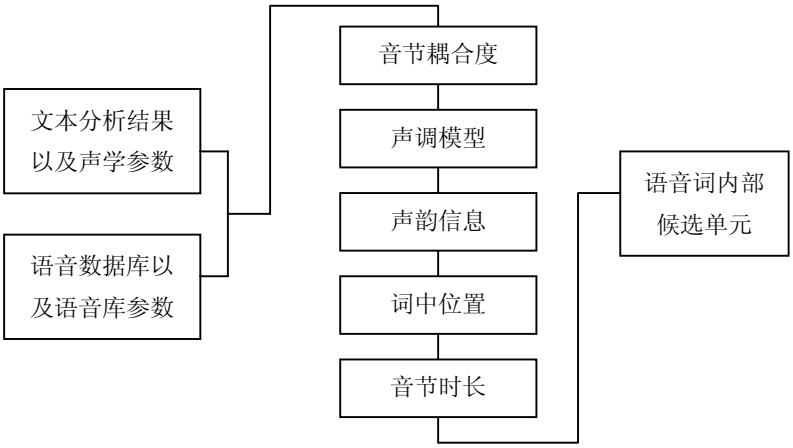

图 2：语音词内部基元选取基本流程

通过这种方法，减少了基元选取过程中所需要考虑的参数的数目，并且通过分门别类，大大简化了基元选取的过程，减小了计算复杂度，避免了很繁重的计算。

(2) 语音词的精选

使用上述方法进行语音词内部的基元选择，确定了各个语音词内部的候选单元。一般来说一个语音词可能存在一个或者多个候选单元。

另一方面，由于上述候选单元的选取完全是局限在单独的语音词内部，不同语音词的候选单元之间完全是孤立的。如果以此而直接将各个语音词的候选单元“生拼硬接”，其效果显然是不理想的。

这就需要使用各个语音词之间的信息对多个候选单元进行精选，而这些信息则主要是文本和语法分析阶段产生的语言学和韵律学信息。其考虑的参数主要包括：句中位置、语音词间声调模型、语音词间韵律属性、重音属性以及语句参数等。

经过语音词内部基元的选取以后，对每个语音词形成一个候选单元的列表，然后对于各个候选单元，通过词语之间的参数选取，从而最终确定每个语音词的基元。

(3) 效率的考虑和实验

人们在评价一个系统的时候，往往从效果（可懂度，自然度）和效率等方面来进行考虑。而从系统的效率方面来说，一个实用系统的基本要求是其实时性。对于不同的方法，其效果和效率有很大的差别。实验中，我们进行了多方面的研究和比较：

a. 完全匹配方法

   听觉效果比较好，但是运行速度慢，不可能实现实时的系统。

b. 基于语音词的滑动窗口匹配算法

   在保持有效的听觉效果的前提下，能够有效的提高系统的效率，加快系统合成的速度，基本上实现实时的要求。但是，在遇到个别候选单元比较多的合成单元时其速度依然会明显降低，效果依然不佳。

c. 提高搜索效率的考虑和实验

   音库搜索程序的优化

音库的重新组织
特征参数集的简化

五、实验结果

我们已实现了一个基于语音数据库的汉语 TTS 系统。在实现过程中，我们研究了汉语的韵律模型、基元的选取等问题。新系统的性能具有较大改善。目前，模型和算法都在不断的改进和优化中。

参考文献

[1] 胡明扬，说“词语”，《语言文字应用》，1999 年第 3 期
[2] 应宏等，基于结构助词驱动的韵律短语界定的研究，《中文信息学报》，1999.6
[3] 王洪君，《汉语非线性音系学》，北京：北京大学出版社，1999 年
[4] 冯胜利，《汉语的韵律、词法与句法》，北京：北京大学出版社，1997 年
[5] 陶建华等，汉语 TTS 系统中可训练韵律模型的研究，《声学学报》