文章编号:1002-8684(2005)10-0057-03

情感语音合成的研究

·论文·

周 洁[1,2], 赵 力[2], 邹采荣[2]
(1. 云南交通职业技术学院 基础部, 云南 昆明 650101; 2. 东南大学 无线电工程系, 江苏 南京 210096)

【摘 要】介绍了语音信号中的情感语音合成的方法,通过分析情感语句的语调,得到了喜、怒、惊、悲4种情感不同的变调规律,对不同的情感类型确定相应的基音频率变化规律、能量变化规律、元音的变异规律和无声时延比例变化规则。对于待合成的语音,首先进行文本扫描,再叠加相应情感的语调变化规则,利用PSOLA算法进行情感语音合成,获得了较好效果。

【关键词】语音信号; 情感语音合成; PSOLA算法
【中图分类号】TN912.34 【文献标识码】A

Study on Emotional Speech Synthesis

ZHOU Jie[1,2], ZHAO Li[2], ZOU Cai-rong[2]
(1. Yunnan Traffic Professional Technology Academy, Kunming 650101, China;
2. Department of Radio Engineering, Southeast University, Nanjing 210096, China)

【Abstract】 Method to generate emotion in synthesized speech is introduced in this paper. By analyzing the tonal rule in emotional speech,the mutative tonal rule was found. Pitch rule, energy rule, vowel rule and silence rule are discovered for corresponding emotional speech. For a text to be synthesized,text scan should be done at first, and then the tonal rules are applied to them. A synthesis emotional speech is finished after that the PSOLA based on tonal rules is done. Experiment based on it shows better performance than ever.

【Key words】 speech signal; emotional speech synthesis; PSOLA

1 引言

语音是人类交际的最重要的工具之一。语音信号处理作为一个重要的研究领域至今已有几十年历史了。人类的说话中不仅包含了文字符号信息,而且还包含了人们的感情和情绪的变化。然而在传统的语音信号处理中往往忽略了包含在语音信号中的情感和情绪因素。在现代语音信号处理中,分析和处理语音信号中的情感特征, 判断和模拟说话人的喜怒哀乐等是一项意义重大的研究课题[1]。

由人工制作出语音称为语音合成(Speech Synthesis)。语音合成是人机语音通信的一个重要组成部分。语音合成技术分为3大类:参数合成方法、波形编辑合成方法和规则合成方法[1]。参数合成技术的算法复杂,并且在压缩比较大时,信息丢失亦大,合成出的语音总是不够自然、清晰。而波形编辑技术用于语音合成时,不存在参数提取的问题, 它通过选取音库中采取自然语言的合成单元的波形, 对这些波形进行编辑拼接后输出。规则合成法是一种高级合成方法。它通过语音学规则产生语音,可以合成无限词汇的语句。情感语音的合成属于语音的规则合成(Synthesis-by-Rule),这里包含2个方面:其一是合成技术的选择,其二是合成规则的制定。

情感发音的实现, 需要通过语音的声学参数体现人的情感特性, 在语调方法的基础上初步加入情感控制参数,增加了语音合成的表现力。吴禀雅[2]等人通过分析不同文本类型的语体色彩和感情色彩, 也较好地改进了合成语音的自然度。Cohn[3]针对情感的声学特性编写了简单的情感编辑器, 使研究人员可细致地观测情感控制参数对语音输出的影响, 对情感语音合成的研究起到了较好的推动作用。笔者介绍了利用PSOLA算法进行情感语音的合成及用PSOLA算法实现情感语音合成时的3个步骤,即基音同步分析、基音同步修改、基音同步合成。研究了喜、怒、惊、悲4种情感语句的变调规律。

【基金项目】国家教育部博士点基金、教育部科学技术重点项目(03082).
国家自然科学基金(60472058).

2 PSOLA 算法合成情感语音

20 世纪 80 年代末，F. Charpentier 和 E. Moulines 等提出了基音同步叠加技术(PSOLA)[4]，它既能保持原始语音的主要音段特征，又能在语音拼接时灵活调整其基音、能量和音长等韵律特征，因而很适合于汉语语音的规则合成[1]，同时汉语情感语音的词调模式、句调模式都很复杂，在以音节为基元合成语音时，单音节在句子中声调、音强和音长等参数都要按规则进行调整[5]。

PSOLA 是用于波形编辑合成语音技术中对合成语音的韵律进行修改的一种算法[6]。决定语音波形韵律的主要时域参数包括音长、音强、音高等。音长的调节对于稳定的波形段来说较简单，只需以基音周期为单位加/减即可。但对于语音基元本身的复杂性，实际处理时采用特定的时长缩放法；音强改变只要加强波形即可。但对于一些重音有变化的音节，有可能幅度包络也需改变；音高的大小对应于波形的基音周期。对于大多数通用语言，音高仅代表语气的不同和话者的更替。但汉语的音高曲线构成声调，声调有辨义作用，因此汉语的音高修改比较复杂。

基音同步叠加技术的实现一般有 3 种方式[7]：时域基音同步叠加(TD-PSOLA)、线性预测基音同步叠加(LPC-PSOLA)和频域基音同步叠加(FD-PSOLA)。笔者采用时域基音同步叠加法进行情感语音合成。时域基音同步叠加技术作为基音同步叠加技术的一种，通过以下步骤实现情感语音的合成：

①对情感语音合成单元设置基音同步标记。同步标记是与合成单元浊音段的基音保持同步的一系列位置点，它们必须能准确反映各基音周期的起始位置。PSOLA 技术中，短时信号的截取和叠加，时间长度的选择均是依据同步标记进行的。浊音有基音周期，而清音的波形接近于白噪声，所以在对浊音信号进行基音标注的同时，为保证算法的一致性，可令清音的基音周期为一常数。

②以情感语音合成单元的同步标记为中心，选择适当长度(一般取两倍的基音周期)的时间窗(如汉宁窗)对合成单元做加窗处理，获得一组短时信号。

③在情感语音合成规则的指导下，调整步骤①中获得的同步标记，产生新的基音同步标记。通过对合成单元标记间隔的增加、减小来改变情感合成语音的基频；通过幅度的变化来改变合成语音的能量；通过对合成单元同步标记的插入、删除来改变合成语音的时长；通过插入无声段，来改变无声比等手段。根据相应规则，采用上述手段，得到符合要求的情感。

④根据步骤③得到的情感合成语音的同步标记，对步骤②中得到的短时信号进行叠加，从而获得情感合成语音。

概括起来，用 PSOLA 法实现情感语音合成时主要有 3 个步骤，即基音同步分析、基音同步修改、基音同步合成。

3 情感语句中的变调规律

笔者用情感语音数据库中的喜怒惊悲 4 种情感的语句各 200 条，研究分析这些情感语句的变调规律。经过分析，发现这 4 种情感分别具有如下的变调规律[9]。

(1)喜。含喜的语句的时长和平叙句相当，但这主要是由句子的尾部带来的影响，句子的前部和中部都比相应内容的平叙句的语速要快一些。句子的振幅强度也集中在句子末尾的一两个字，整个句子的声调的调域要比平叙句高。由于句子的前中部语速加快，受到生理原因和语法条件的制约，句中非关键性的字和词的调形拱度就变得平坦一些，甚至失去本调，而成为前后相邻两调的中间过渡。句尾的感叹词在平叙句中读轻声，在这里语气有很强的加重，并且调形变成为先升后降的山包形。

(2)怒。含怒的语句的时长约为平叙句的一半左右，其振幅强度也很高，是加速句和加强句的结合。句中的动词和修饰动词的副词其振幅强度比平均值要高一些。句子的调域抬高，但调形不一定变平，有时它们的拱度甚至更加扩展了。句尾的感叹词也不同于轻声，而变成类似于上声的声调。

(3)惊。含惊的语句的情况和含喜的语句相类似，不同之处在于句尾的调形有上翘的趋势。整个句子的平均振幅强度比平叙句略高，原因在于句尾的振幅强度增高了。

(4)悲。含悲的语句的时长约为平叙句的一倍左右，其振幅强度也低许多。由于每个字的读音彼此都拉得很开，所以字调的调形保留了其单字的调形，多字调的效果弱化了。但由于悲的语句中几乎每个字都夹杂了一定程度的鼻音，所以要进行鼻化的处理。含悲的语句调域降低，整个语句趋于平坦化。

根据对情感语句特征参数的考察以及听音者的主观感觉，笔者改变某一普通语句的局部调形(拱度)，或使其整句的调域有所改变，使其能够反映响应的情感语意。在反映不同情感的语句中，各基本单元的调形基本上稳定，但它会产生一些调位变体。

通过分析可以看出，情感的变化在语音中主要表现为几个方面。(1)基音频率的变化。这主要体现为不同情感下基频的偏移。比如高兴时说话人的基频就会较一般状态下高。(2)能量的变化。这主要体现为高激活情感状态下能量的增加，某些特定情感对应的特定情绪词的重读等现象。(3)元音的变异。主要体现元音的延时和模糊化。(4)无音间隙的插入。主要体现为通过无声时间的比例来表现情感的特征。因为在语言交际中，说话人的习惯不同，语言环境不同，就不可能有什么铁定不变的规律。但其基本趋势是不变的。图1是情感语句“下雨了(啦)！”这句话的4种情感的变调情况[9]。

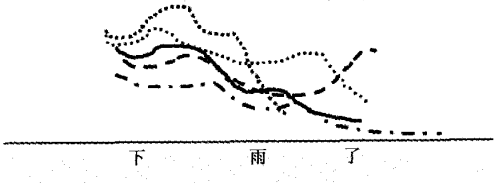

图1 情感语句变调实例

获取同类情感，分析总结不同情感下所对应的基音频率变化规律、能量变化规律、元音的变异规律和无声时延比例，总结规律，将其作为进一步进行PSOLA合成的规则使用。

4 情感语句合成系统

整个合成系统可用软件实现，软件分为3个模块：文本扫描模块、语音合成模块以及放音模块。在文本扫描模块中，对输入的汉语拼音、调型、空格、标点符号等组成的文本进行分词、分字处理，分析出一句文本中的音素表、词表、句表以及控制符和停顿等信息。

语音合成模块是整个情感语音合成系统中最重要的组成部分，它的核心包括PSOLA算法和韵律调节2个部分。其工作过程为：根据上面得到的一句文本信息，调用韵律规则库，决定选用一组合适的词调模式、句调模式、音长模式和音强模式，从而计算出适合于该句的句调和词调曲线，以及能量的分配关系。然后由音素表从全音节音库中取出相应的音节数据及该语音的基音同步标记和能量等相关信息，再对这些语音进行韵律、音强和音长的调整。在调整过程中根据文本的控制信息插入适当停顿，从而最终得到一句流畅、自然的汉语情感语音，并将其送入放音模块的数据缓冲器中。最后由放音模块将放在缓冲器中的合成语音数据通过声霸卡上的D/A转换和扬声器播出。

整个系统包括音库、韵律库以及情感语音合成软件，均遵循模块化的设计原则，便于系统维护、规则的完善及功能扩展。

5 情感语音合成实验

笔者采用了4个句子进行情感语音合成实验。句子由单字词和二字词组成。被测试者为5人，让被测试者对情感语音合成后的句子的可懂度和自然度进行评价。也将实验结果自然度分为优、良、中、差、极差5个等级。实验的4个句子分别为“下雨了”、“你真伟大呀”、“快点干”和“这下全完了”。

实验中被测试者可以全部听懂句子的意思，说明用笔者采用的合成和韵律控制方法，合成语音的可懂度是比较令人满意的。在5个被测试者中，有2人认为总体自然度效果为优，2人认为总体自然度效果为良，1人认为总体自然度效果为中。图2为“下雨了”的合成语音波形。

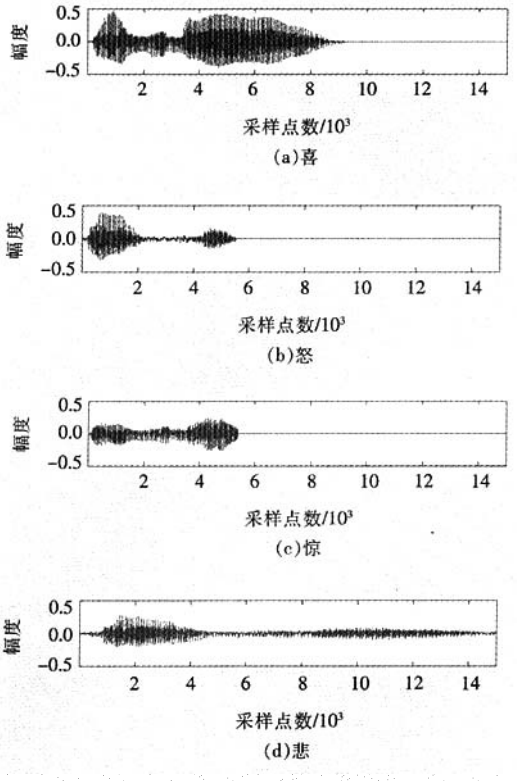

图2 “下雨了”在喜、怒、惊、悲4种情感下的合成语音波形图

(下转第73页)

(1)硬盘阵列

由于硬盘具有读写速度快、在线检索、并发读写等特点,所以可用来存储大量常用的数字化资料,并且可以满足多人的在线并发访问。尤其是目前广泛使用的光纤硬盘阵列,不仅传输速率高且稳定可靠,随着硬盘阵列存储技术的不断进步,硬盘容量的不断提高,硬盘阵列技术的应用将会越来越广泛。

(2)光盘和光盘库

光盘作为一种新的存储媒体,感光层上有一层保护层,所以适合长时间保存,上面的数据不会丢失。另外,光盘数据读取为非接触性激光方式,为无检索磨损。随着光盘存储技术的飞速发展,光盘存储容量有了极大的提高。光盘体积小,能大大降低存储空间从而降低存储成本。目前,普通的 CD-R 容量为 650 MB,价格便宜,可以作为 WAV、S48 等众多文件的存储媒介。一个光盘库能存储几十片或几百片的光盘,光盘适于几个人的后备查询,它利用机械手可实现自动换盘,并可以同时对多张光盘进行读写操作。用专门的光盘库管理软件,允许用户通过计算机网络直接进入光盘库查询检索文件素材,还可通过网络对光盘库进行管理,如权限设置、状态监视、读写控制等。

(3)数据流磁带库

在存储介质中,除了硬盘阵列和光盘库,还可采用数据流磁带库,它是一种相对廉价的存储方式,一盘与普通录像带同样大小的数据流磁带上可以存储容量为几百 GB 的节目素材,并且读取速度可达到 12 MB/s,由于数据流磁带采用顺序读写方式,当用户在线访问时,要先将磁带上的数据迁移到服务器的缓存中,需要一定的时间,所以它只适合用于素材的备份,而不适用在线访问。

7 结束语

广播设备数字化后带来的一些新问题、新要求,是数字化进程中的必然结果,随着科技的发展,这些问题将会迎刃而解,但新的问题可能又将产生,列举这些要求,并作出相应的应对措施,目的是为了让大家在使用数字设备时,充分了解其特性,以便使用好数字设备,充分发挥其效能。更深层次的是要改变观念和工作模式,树立系统概念,从以往的以物使用逐步变成按需使用,这将是以后数字设备的应用前景。

参考文献

[1] Ken C. Pohlmann 著. 数字音频原理与应用. (第 4 版)苏菲译. 北京:电子工业出版社, 2002.

[2] 数字演播室的电缆应用 TB-65. 美国百通(Belden)电线电缆网. http://www.belden.com.cn/.

[3] AES/EBU 规格. 日本佳耐美(Canare)电子有限公司. http://www.canare.com.cn/

[收稿日期] 2005-05-13

(上接第 59 页)

6 结论

分析了喜、怒、惊、悲 4 种基本情感语句中所包含的变调规律。获取不同情感下所对应的基音频率变化规律、能量变化规律、元音的变异规律和无声时延比例,将其作为进行 PSOLA 合成的规则使用。用 PSOLA 算法从基音同步分析、基音同步修改、基音同步合成等方面分析了情感语音的合成。整个合成系统由文本扫描模块、语音合成模块以及放音模块实现。实验表明,基于情感语调规则的 PSOLA 方法能对情感语音合成能取得较好效果。

参考文献

[1] 赵力. 语音信号处理. 北京:机械工业出版社,2003.

[2] 吴禀雅,周昌乐,吴洁敏. 汉语基调的调模与语音合成的质量提高. 中文信息学报, 2003,17(3):23-28.

[3] Cahn J. Generating Expression in Synthesized Speech. Master's thesis, MIT, 1989.

[4] 姚添任. 数字语音处理. 武汉:华中理工大学出版社, 1992.

[5] 陈永彬,王仁华. 语音信号处理. 合肥:中国科学技术大学出版社,1990.

[6] 赵力,小林丰,新美康永. 音素情报と音调情报を统合した中国语连续音声认识. 信学技术,1996,SP96(12):121-128.

[7] 涂湘华,蔡莲红. 用于语音合成的 PSOLA 算法简介. 微型计算机,1996,16(5):5-9,23-28.

[8] 吴宗济. 普通话语句中的声调变化.中国语文,1982,(6):439-449.

[9] 赵力,钱向民,邹采荣. 语音信号中的情感识别研究. 软件学报,2001,12(7):1 050-1 055.

作者简介

周洁,硕士研究生,研究方向为语音信号处理.

赵力,教授、博士生导师,研究方向为语音信号处理.

邹采荣,教授、博士生导师,副校长,研究方向为信号处理.

[收稿日期] 2005-04-17