·学术探讨·

合成语音自然度客观测度

赵　博　蔡莲红
（清华大学计算机系人机交互与媒体集成研究所，北京 100084）

摘　要　目前合成语音的自然度有待提高，论文根据目前的研究现状提出了一种合成语音自然度的客观评价方法，该方法主要从语音韵律特征的主要参数出发，计算同一发音人的自然语音和合成语音之间的基频、时长、音强等参数的差距，其中由于两种语音基频时间不匹配，所以采用 DTW（Dynamic Time Warping）算法来对两种语音的基频进行了时间弯折对准。最后再将计算结果与主观评测（MOS）的结果进行比较。实验数据表明，论文提出的基频曲线失真测度与 MOS 之间具有很强的相关性，从韵律特征角度给出的评价结果能够衡量合成语音的自然度。

关键词　语音合成　评测　自然度

文章编号 1002-8331-(2005)07-0032-02　**文献标识码** A　**中图分类号** TP37

Objective Measure of Naturalness for Concatenate Speech Synthesis

Zhao Bo　Cai Lianhong
(Institute of Human-Computer Interaction and Media Integration, Tsinghua University, Beijing 100084)

Abstract: In this paper, a new objective evaluation method of naturalness for concatenate speech synthesis is proposed. Considering the prosodic parameters of speech, the objective distance of patch parameters, duration parameters and intensity parameters between the natural speech and the synthesized speech are calculated. For mismatch of two speeches in duration, the DTW (Dynamic Time Warping) algorithm is used to allow approximate matching. The formal Mean Opinion Score (MOS) obtained subjectively is compared with the result. The correlation coefficient between the objective measure and subjective measure is strong. The experiments show that the proposed method can serve as the objective evaluation of naturalness for concatenate speech synthesis.

Keywords: speech synthesis, evaluation, naturalness

1　介绍

合成语音的自然度还有待提高，因此对合成语音自然度的评测也有深入研究的必要，目前合成语音自然度的评测普遍采用的是主观评测方法，主要采用的方法有主观印象分 MOS 和两两对比测试 PC 等，其优点是符合人对语音自然度的感觉，缺点是费时费力、灵活性不足，而且稳定性和重复性较差，受人的主观影响比较大。为此国内外有人提出了对合成语音进行客观评测的方法，国外对合成语音自然度的客观评测方法主要有：根据分段覆盖所导致的拼接代价评测[1]和根据语音参数距离评测[2]等，而国内采用的则有根据人耳听觉特性进行评测[3]等方法，初敏等人对韵律与自然度的关系进行了研究[4]。现在的语音合成系统大多数都采用了基于大规模语音数据库的波形拼接来产生合成语音，用来拼接的语音基元来源于包含大量语句的自然语音数据库，由于这些语音基元来自自然语音，因此有较高的清晰度，但是这些语音基元往往还体现了其所在上下文的音段和韵律特性，所以在自然度上还有所欠缺。此类语音合成系统的关键就在于语音基元选取的算法上要考虑与语音相关的音段和韵律特征。研究表明韵律特征主要体现在基频、时长、幅度等声学参数上，而且相邻基元的韵律特征的变化是影响语音自然度的重要因素[5]。

基于目前的语音合成发展现状，为了帮助语音基元选取算法的研究和发展，论文提出了通过计算同一发音人的自然语音和合成语音之间的韵律特征参数距离，来对合成语音自然度进行客观评测的方法。

2　语音的韵律参数提取

连续的自然语句中，同一音节的发音会受到上下文的影响而产生各种变化，这些变化也体现出了某一特定音节因为语句的韵律不同而引起的超音段特征（如基频、时长、幅度等韵律特征）方面的变化。目前由于大规模语音数据库来源于播音员，因此单音的清晰度和自然度都比较高，然而同样因为语音来自自然句录音，每个单音中还包含有超音段信息，在拼接合成时这些超音段信息的不匹配影响了语音的自然度，由此语音合成技术的主要问题就集中在基元选取算法方面，而如何提取合理的参数以提高合成语音的自然度就成了目前的研究重点，这当中影响较大的就是韵律特征信息。

对于发音相同、语境不同的音节，由于在音段内容方面保留了绝大部分相同的特征，相比之下，它们在超音段内容方面的变化要显著得多，因此超音段内容的变化是决定音节知觉差异的最主要因素。而超音段内容的变化主要通过基频、时长和

基金项目：国家自然科学基金项目（编号：60275014）
作者简介：赵博（1974-），男，硕士研究生，主要研究方向为语音合成系统评测。蔡莲红，女，教授，主要研究方向为多媒体技术和语音合成。

音强三个声学参数体现出来[6]。同时吕士楠等的研究也表明合成语音与自然语音在时长和停顿上有较大差异[7]。论文就是从韵律特征参数出发分析自然语音和合成语音的韵律参数距离，同时考虑了时长和停顿这些节奏方面的因素。其中主要采用了语音基频、语音时长、音强和静音段时长等参数。在这些参数中最重要的就是语音基频，语音基频曲线中包含有较多的超音段信息，而比较自然语音和合成语音之间的基频曲线距离，就能够体现出合成语音在自然度方面的差距。

其中语音基频和时长都为自动标注并由人工修正，语音波形和基频曲线如图 1 所示，音强采用归一化的平均幅度。

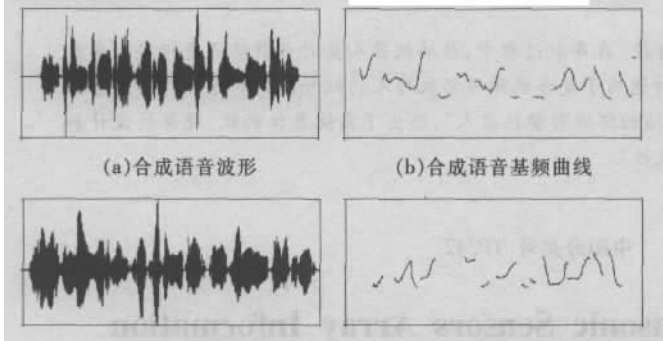

图 1　合成语音和自然语音比较

对于基频特征参数的处理比较特殊，对于一个特定音节来说，两种语音的基频都是由一个序列构成的，而且序列长度一般不会相同，对此就采用了 DTW 算法对其进行时间对准处理。由 DTW 算法计算出来的距离为两种语音基频序列的欧氏距离的平方，还要根据两个序列的长度来计算两个音节基频的距离。对一个语句中的所有音节的基频距离进行算术平均，得到整个语句的基频距离。计算公式如下：

$$D_p=\frac{1}{K}\sum\sqrt{\left(\frac{DTW(P_s,P_n)}{(M\cdot N)^{1/2}}\right)} \tag{1}$$

其中，K 为语句中包含的音节个数，M 和 N 分别为语句中对应于某音节的基频序列长度，P_s 和 P_n 分别为对应的某音节的合成语音和自然语音的基频序列。

对于音节时长和静音段时长，只要简单地计算对应时长的均方根即可，为了处理方便以自然语音的时长为单位作归一化处理即可。其计算公式如下：

$$D_t=\sqrt{\frac{1}{KT_a}\sum(T_s-T_n)^2} \tag{2}$$

其中，K 为语句中包含的音节个数或者静音段数，T_s 和 T_n 分别为对应的合成语音和自然语音的时长，T_a 为 K 个音节时长的平均值。

对于音强的参数距离，则计算对应音节归一化平均幅度的均方根，其计算公式如下：

$$D_e=\sqrt{\frac{1}{KE_a}\sum(E_s-E_n)^2} \tag{3}$$

其中，K 为语句中包含的音节个数，E_s 和 E_n 分别为对应音节的合成语音和自然语音的音强，E_a 为音强的平均值。

3　合成语音和自然语音参数距离计算

为了从两种语音的特征参数得出对合成语音的自然度评测结果，这里采用了合成语音和自然语音的参数距离来计算两种语音的自然度差距，或者说两种语音的参数距离，由下式表示：

$$D=\sum w_iD_i,\ \sum w_i=1 \tag{4}$$

其中，D_i 表示第 i 个参数的距离，而 w_i 则表示第 i 个参数距离在自然度评测中的权重。由前节所述，采用的 4 个参数分别是：基频、音节时长、音强和静音段时长。这 4 个参数距离各自的计算方法各不相同。对于两种语音的基频来自人工修正的自动标注，由于合成语音也是由经过标注的自然语音拼接处理产生的，因此可以通过程序自动产生其标注信息。音节时长和静音段时长的处理与基频处理方法类似，都可以通过程序来产生，而音强则是采用归一化的平均幅度来表示。

4　与主观评测 MOS 的实验比较

客观评价和主观评价之间的关系常用一种函数映射关系来表示，这里采用二次多项式拟合。其公式表达为：

$$M'=a+bO+cO^2 \tag{5}$$

其中 M' 表示由客观评价方法预测出的主观评价值，O 为客观评测值，a、b、c 为二次多项式系数。由于客观评价实质上是对主观评价值的一种预测，所以客观评价方法的性能好坏是以其与实际主观评价值的相关性来衡量的。两种评测方法的相关性可用它们的相关度 ρ 和标准差 σ 来表示。计算公式如下：

$$\rho=\sqrt{\frac{\sum_{i=1}^{N}(M'-\mu_m)^2}{\sum_{i=1}^{N}(M-\mu_m)^2}},\ \sigma=\sigma_m\sqrt{1-\rho^2} \tag{6}$$

其中，N 为样本数，M 表示实际主观评价值，μ_m 为实际主观评测的均值，σ_m 为实际主观评价值标准偏差。

在实验中采用了 20 句合成语音和 20 句相同文本内容的自然语音，由 7 名听音人按照 7 分制 MOS 给出其认可的自然度，7 个分制分别为：5 非常自然，4.5 自然，4 比较自然，3.5 不太自然，3 可接受，2 比较差，1 不能接受。根据前述的公式分别计算合成语音和自然语音的几个参数距离，利用线性加权和法（公式(4)）构成总的参数距离，并用最优化方法解出对应的权重 w_i，使得用这种方法获得的客观评测结果与 MOS 得到的自然度得分相关度达到最大，解得 w_1=0.9255，w_2=−0.9315，w_3=0.6742，w_4=0.2915。并得到公式(5)中的三个系数分别为 a=3.7716，b=−5.1837，c=5.7342。而客观评测结果与 MOS 得分的相关度和标准差为：ρ=0.8152，σ=0.1623。另外不考虑时长和幅度的情况下，仅通过基频差距获得的客观评测结果与 MOS 得分的相关度为 ρ=0.7462。

5　结论

通过前面的实验和研究表明，两种语音的基频差距与主观评测获得的自然度有比较高的相关度为 0.7462，而再加上时长、音强等参数通过线性加权法获得的联合评测结果，与主观 MOS 评测结果有更高的相关度，达到了 0.8152。从这个结果来看，通过计算合成语音与自然语音的韵律参数距离，可以获得合成语音的自然度评测。同时这个结果也说明了基元的韵律特征等超音段特征与合成语音的自然度有比较大的相关性。

（收稿日期：2004 年 11 月）

（下转 152 页）

(5)网络地址转换(NAT):NAT 容许一个单独的设备(例如路由器)在 Internet(公共网)和 Intranet 本地(私有)网之间充当代理,这就说明只用一个 IP 地址就可以代表整个组的计算机,从而为解决 Internet 地址耗尽和路由表爆炸等危机提供可行的方法。路由器通过分析公共网的报文头来分类报文,所用的分类规则就是内外网对应的地址、端口转换表,可能用到的分类字段有网络源、目的地址,源、目的端口号等。规则不需频繁更新。

(6)网络计费:网络计费是网络管理的重要组成部分,开发一套完善的网络计费系统是每个网络运营部门的首要工作。网络计费主要任务是对登录用户进行认证和传输测量(例如测量两个子网间的传输量),目前网络普遍采用的是以 IP 地址标识用户,统计该 IP 地址的数据流通量的计费方法(可能按目的地址、应用等分类统计)。规则多为 2 维或 2 维以上(不需频繁更新),包括:源、目的 IP 地址,源、目的端口等。

(7)负载均衡(load balance):负载均衡技术常根据源、目的 IP 地址或应用层协议的不同,将较大的网络流量按预先制定的策略分配到一系列服务器上。这种应用的规则不需要频繁更新。

除了上面介绍的应用外,报文分类在网络测量、拥塞控制、资源预留、收集网络统计数据、路由器和交换机的设计、第 4 层交换、MPLS(multiprotocol label switching)通道的流合并等应用中都有广泛的应用。在已知应用对分类技术需求的情况下,可以根据分类算法的适用维数、更新复杂度、时/空复杂度等因素合理地选择适当的分类算法,也可以采用多种分类算法的组合。

表 2 常见分类算法及其性能分析

算法	最坏时间复杂度	最坏空间复杂度	更新的复杂度	(未经改进的)算法适用的维数
Linear search	N	N		d
Hierarchical Tries	W^d	NdW	增加 d^2W	d
Set-pruning Tries	dW	N^d	N^d	d
Grid-of-Tries	W^{d-1}	NdW	NW	2
Cross-producting	dW	N^d	更新需要重新构建数据结构	d
AQT	αW	NW	$\alpha\sqrt[\alpha]{N}$	2
FIS-tree	$(L+1)W$	$L\times N^{L+1/L}$		2
RFC	d	N^d		d
HiCuts	d	N^d		d
Tuple Space Search	N	N	1 次 Hash 操作	d
Ternary CAM	1	N		d
Bitmap-intersection	$dW+N/memwidth$	dN^2		d

6 总结与展望

随着网络带宽的增加和用户对于网络服务多样性要求的增长,具有报文分类功能的网络设备将会不断涌现,而报文分类算法也将成为研究的热点。目前的报文分类算法大多受到处理速度、存储空间、规则维数(d)、长度(W)、更新难度、可扩展性和移植性等等因素的影响难以得到广泛的应用。

未来的报文分类算法一方面将追求较高的速率,这势必需要采用高性能硬件平台(例如网络处理器等),针对不同的平台研发高性能报文分类算法将是其发展的趋势之一;另一方面,未来的报文分类算法将更加追求可扩展性和可移植性,尽量做到在较小的硬件(内存、处理速度等)需求下提高处理速度,并且使得算法尽量不受或少受规则个数 N、维数 d 和匹配长度 W 等参数的影响。

同时分类问题是个广泛的问题,不仅仅出现在报文分类领域,还在包括 DNA 匹配、物种分类、货物装载(集装箱装载)等领域得到了广泛的应用。而在这些领域已经广泛采用了的算法可以和报文分类问题互相借鉴、参考。(收稿日期:2004 年 10 月)

参考文献

1.V Srinivasan,S Suri,G Varghese et al.Fast and Scalable Layer four Switching[C].In:Proceedings of ACM Sigcomm,1998-09:203~14

2.M M Buddhikot,S Suri,M Waldvogel.Space decomposition techniques for fast layer-4 switching[C].In:Proceedings of Conference on Protocols for High Speed Networks,1999-08:25~41

3.A Feldman,S Muthukrishnan.Tradeoffs for packet classification[C].In:Proceedings of Infocom,2000-03;3:1193~202

4.Pankaj Gupta,Nick McKeown.Packet Classification on Multiple Fields[C].In:Proc Sigcomm,Computer Communication Review,Harvard University,1999;29(14):147~60

5.Pankaj Gupta,Nick McKeown.Packet Classification using Hierarchical Intelligent Cuttings[C].In:Proc Hot Interconnects VII,Stanford,1999-08,IEEE Micro,2000;20(1):34~41

6.T V Lakshman,D Stiliadis.High-Speed Policy-based Packet Forwarding Using Efficient Multi-dimensional Range Matching[C].In:Proceedings of ACM Sigcomm,1998-09:191~202

7.P Tsuchiya.A search algorithm for table entries with non-contiguous wildcarding[R].unpublished report,Bellcore

8.V Srinivasan,S Suri,G Varghese.Packet Classification using Tuple Space Search[C].In:Proceedings of ACM Sigcomm,1999-09:135~46

9.Pankaj Gupta,Nick McKeown.Algorithms for Packet Classification.Computer Systems Laboratory,Stanford University Stanford,2001

10.田立勤,林闯.报文分类技术的研究及其应用[J].计算机研究与发展,2003;(6)

11.喻中超,徐恪,吴建平.一种适用于多维的快速 IP 分类算法[J].软件学报,2001;(12)

(上接 33 页)

参考文献

1.Robert Bat ušek.An Objective Measure for Assessment of the Concatenative TTS Segment Inventories[J].Eurospeech,2001

2.Jun Xu,Cuntai Guan,Haizhou Li.An Objective Measure for Assessment of a Corpus-Based Text-to-Speech System[C].In:IEEE 2002 TTS Workshop,2002

3.陈静,周毅刚,周建林.符合人耳听觉特性的语音音质的客观评价方法[J].哈尔滨工业大学学报,1998-06

4.初敏.韵律研究与合成语音的自然度[C].见:第五届全国现代语音学学术会议-新世纪的现代语音学,北京:清华大学出版社,2001:295~301

5.吴志勇,蔡莲红.语音合成中的韵律关联模型[J].中文信息学报,2004;18(2):44~50

6.周讯溢,王蓓,杨玉芳等.语句中协同发音对音节知觉的影响[J].心理学报,2003;35(3):340~344

7.吕士楠,林凡,张连毅.基于大语音库的拼接合成语音特征分析[C].见:第五届全国现代语音学学术会议——新世纪的现代语音学,北京:清华大学出版社,2001:307~310