MATLAB 在语音分析中的应用①

赵 博 （清华大学计算机系人机交互与媒体集成研究所 100084）

摘要：MATLAB 作为一种科学计算工具，在科学研究的各个领域得到了广泛的应用。在国家 863 计划语音合成系统评测中，较多使用的还是听音人参与的主观方法，我们利用 MATLAB 工具实现了对语音质量进行客观评测的方法，本文通过阐述这一方法的实现过程，描述了 MATLAB 在语音分析方面的一个具体应用。

关键词：MATLAB 语音 信号处理

1 引言

MATLAB 是一个包括数值计算，高级图形和可视化的集成科技计算环境，它也是一种高级程序设计语言。不管你有什么问题－算法，分析，图形，报告或者模拟－MATLAB 可以帮助你。灵活的 MATLAB 语言可使工程师和科学家简练地表达他们的思想。强有力的数值计算方法和图形便于测试和探索新的思想，而集成的计算环境便于产生快速的实时结果。MATLAB 得到了各个领域专家学者的广泛关注，其强大的扩展功能为用户提供了强有力的支持；它集数学计算、图形计算、语言设计和神经网络等 30 多个工具箱于一体，具有极高的编程效率。

由于采用了大规模自然语音资源库进行波形拼接合成，现在语音合成系统所产生的合成语音具有很高的清晰度和可懂度，但是合成语音的自然度还有待提高。因此需要对合成语音的自然度进行评测以利于语音合成系统的发展，目前普遍采用的还是由听音人进行试听，再根据主观印象进行打分评比的方法。本文作为对比也采用了 7 分制主观印象打分（MOS），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。

另一方面随着计算机系统的利用和语音合成技术的发展，使用计算机系统建立起来的基于大规模语音库的语音合成系统也越来越趋向于复杂化，传统的主观评测方法费时费力且灵活性不足，因此就需要研究利用计算机实现自动对合成语音自然度进行评价的客观方法，这样也可以避免主观评测中很难避免的学习效应。本文采用的方法是对基于同样文本的来自同一发音人的自然语音和合成语音，进行 MFCC 特征参数距离计算来获得客观评价结果。

由于 MATLAB 对语音信号分析处理的能力很强大，我们利用 MATLAB 实现了对语音信号 MFCC 参数的提取和距离计算，并利用其强大的数据分析处理功能，对获得的客观评价结果与主观评价结果（MOS）之间的关系进行了计算，证明了客观评价方法的有效性。

2 提取语音信号的特征参数

为了实现通过计算机自动进行合成语音自然度客观评测的目的，我们就要找出能够计算的合成语音和自然语音之间差距方法，并且这个差距要能够体现出对于人耳来说合成语音质量和自然语音质量的差别。本文采用的方法是从语音信号中提取特征参数，比较合成语音和自然语音的特征参数并计算其距离，通过特征参数距离来描述合成语音自然度与自然语音之间的差距，这样就可以把计算得到的特征参数距离作为合成语音的客观自然度评价结果。

通过研究，人们发现人耳对不同频率的语音具有不同的感知能力，这个感知能力并不是随着频率的增加而线性增加。经过大量的实验，人们根据人耳在不同频率下的音调感知能力，提出了 Mel 频率的概念，这里的 Mel 就是人耳所感知到音调的度量单位。由于汉语是有调语言，Mel 频率正是对人耳所听到的汉语音调的度量。通过 MATLAB 工具计算这个参数将可以很

① 项目背景：国家 863 计划语音合成系统评测

好的描述人耳对汉语语音音调的感知情况。很多的研究也证明由于 Mel 频率特性反映了人耳的听觉特性，因而在用于代替人耳来分析语音时，其性能和鲁棒性都是最符合实际听音结果的。

音调和频率之间的关系近似满足方程：$Pmel \equiv (1000/lg2) \times lg(1+0.001fHz)$。Mel 倒谱系数（Mel – Frequency Cepstrum Coefficients，MFCC）就是根据 Mel 频率的概念而提出的，其提取的计算过程如图 1。

图 1　Mel 频率倒谱系数（MFCC）提取过程

由于相对于声波信号，人的发音器官运动速度显得非常慢，所以一般认为人类的语音信号是短时平稳信号，可以对其进行短时分析，最基本的手段就是对语音信号进行分帧，然后再进行分析处理，就是用一个有限长度的窗序列 w(n) 截取一段语音信号来分析。本文采用了哈明窗函数来对语音分帧，每帧长度为 256，步长为 80，计算中利用了 MATLAB 本身的哈明窗函数（hamming）。采用 Matlab 可以很容易的实现对语音信号 s(n) 的分帧。其分帧过程描述如下：

```
function Sn = enframe(s)
% 计算语音分帧后的帧数
nf = fix((length(xx) – 256 + 80)/80)
% 设定分帧的帧长和步长
Sn = zeros(nf,256)
indf = 80 * (0:(nf – 1))'
inds = (1:256)
Sn(:) = s(indf(:, ones(1,256)) + inds(ones(nf,1),:))
% 加入哈明窗(hamming)
for i = 1:nf
Sn = Sn(i,:)'. * hamming(256)
end
```

在 MATLAB 工具的基础上，我们应用了一个语音分析工具箱 VoiceBox，这个工具箱中没有 MFCC 特征参数的直接计算函数，但是包含有 Mel 频率的滤波器系数处理函数 MELBANKM。通过下面描述的算法过程，我们在 MATLAB 工具中实现了 MFCC 参数的提取。除了提取 MFCC 参数外，为了描述语音帧之间的相关性，我们在计算中引入了一阶差分 MFCC 特征参数，并与 MFCC 参数一起构成语音的特征参数。下面是语音信号 MFCC 参数提取的具体实现：

```
function getmfcc = mfcc(s)
% 设定 mel 滤波器系数
bank = melbankm(24,256,8000,0,0.5,'m')
bank = full(bank)
bank = bank/max(bank(:))
% 设定 DCT 系数
for k = 1:12
  n = 0:23
  dct(k,:) = cos((2 * n + 1) * k * pi/(2 * 24))
end
% 设置归一化的倒谱提升窗口
w = 1 + 6 * sin(pi * [1:12]./12)
w = w/max(w)
% 设置预加重滤波器
ss = double(s)
ss = filter([1 –0.9375],1,ss)
% 对语音信号进行分帧
ss = enframe(ss)
% 计算每帧的 MFCC 参数
for i = 1:size(ss,1)
s = ss(i,:)
% 对信号 s 进行 fft 计算
t = abs(fft(s));
t = t.^2;
% 对 fft 参数进行 Mel 滤波取对数再计算倒谱
c1 = dct * log(bank * t(1:129));
c2 = c1. * w';
m(i,:) = c2';
end
```

```
% 计算 mfcc 参数的一阶差分
dtm = zeros(size(m))
for i=3:size(m,1)-2
dtm(i,:) = -2*m(i-2,:) - m(i-1,:) + m(i+1,:) + 2*m(i+2,:)
end
dtm = dtm / 3
% 合并 mfcc 参数和一阶差分参数
ccc = [m dtm]
% 去除首尾两帧,因为这两帧的一阶差分参数为 0
ccc = ccc(3:size(m,1)-2,:)
```

这样我们就可以通过计算 MFCC 参数,获得合成语音和自然语音的特征,由于 MFCC 参数是对人耳听觉特性的描述,因此可以认为这两种语音的 MFCC 参数距离,能够代表人耳对两种语音的听觉上的特性差异,所以可以用这个距离来作为合成语音自然度的客观评价结果。

3 语音特征参数的分析处理

获得自然语音与合成语音的 MFCC 特征参数后,我们还要通过计算两种语音的 MFCC 特征参数距离来获得合成语音自然度的客观评价结果。由于两种语音的时间长度不一致,所以获得的特征参数向量长度也不相同,为了获得最佳的比较效果,我们采用了动态时间规整(DTW)方法来计算两种语音 MFCC 特征向量的距离。

假设两种语音帧长度分别为 N 和 M,动态时间规整算法就是要寻找通过起始点(1,1)到终止点(N,M)的路径,路径上每一个节点(X,Y)的值表示到这个节点为止特征参数的距离 D(X,Y),算法的目的就是要找出从起始点到终止点的最佳路径,使得距离 D(N,M)最小。为了防止盲目搜索,还需要规定搜索范围,在实际应用中我们规定的搜索范围是最大斜率为 2,最小斜率为 1/2,其路径搜索如图 2 所示,其中的细线框表示搜索范围。

在 MATLAB 中没有 DTW 算法的实现工具,因此这一算法我们就需要自己进行设计,由于在这个具体的应用中我们只需要计算出最后的距离 D(N,M),而不需要知道具体的路径,因此计算过程中不需要保存具体的路径信息,算法描述如下:

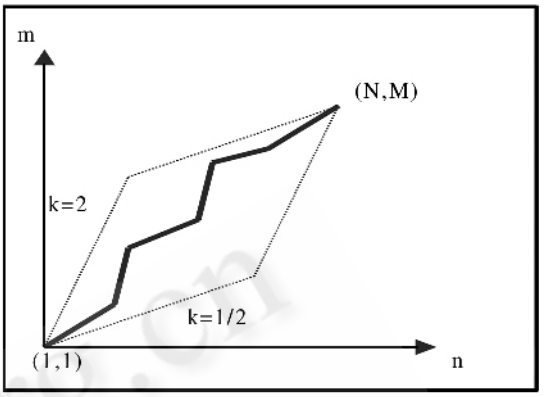

图 2　动态时间规整中的路径搜索

(1) 初始化,设置 n 和 m 的初值分别为两个向量的长度,计算各帧匹配距离矩阵 d(n,m),设定 D(1,1)=d(1,1);

(2) 对于每一个 i=2,3,…..,n,求 j=1,2…..,m 的每一步最小距离 D(i,j):

① 令 D1=D(i-1,j)

② 如果 j>1,令 D2=d(i-1,j-1)

③ 如果 j>2,令 D3=d(i-1,j-2)

d) D(i,j)=d(i,j)+min(D1,D2,D3)

(3) 通过 DTW 得到的最终结果就是 D(n,m)

这个结果可以直接作为合成语音自然度的客观评测结果 dmfcc,为了验证客观评测结果的有效性,这个结果还要与主观评价结果 mos 进行比较,我们需要得到客观评价和主观评价(MOS)之间的关系。

两者间的关系我们可以用相关度和标准偏差来计算,客观评价结果和主观评价结果之间的关系常用一种函数映射关系来表示,这里我们采用二次多项式拟合 dmos=a*dmfcc^2+b*dmfcc+c,通过这个多项式我们可以把客观评测结果 dmfcc 映射到 mos 得分,其中的系数(a,b,c)可以通过 MATLAB 的多项式拟合函数 para = polyfit(dmfcc,mos,2)来得到,我们可以利用 MATLAB 中的数据处理工具来计算相关度 ρ 和标准偏差 σ。其计算方法描述如下:

```
function relat = relation(dmfcc,mos)
% 计算拟合二次多项式的参数
para = polyfit(dmfcc,mos,2)
% 计算相关度和标准差
```

```
p = corrcoef( dmfcc, mos)
o = std( mos,1) * sqrt(1 - p * p)
relat = [para, p, o]
```

其中 ployfit 是 MATLAB 中的多项式拟合函数,其中的参数 2 代表是二次多项式。相关度由 MATLAB 中 corrcoef 相关系数计算获得,再由主观评测 mos 得分的标准差和相关度系数计算出两种评测结果之间的标准偏差。

在实验中,我们采用来自 20 个测试文本的自然语音和合成系统产生的合成语音进行了实验,其中自然语音来自同一发音人的录音,主观评测采用了前面所述的 7 分制 MOS 评分,客观评测则用 MFCC 特征参数距离计算得到。最终我们得到了两种评测结果的相关度 $\rho = 0.8075$ 和标准偏差 $\sigma = 0.1468$。

4 结束语

国家 863 计划多年来对语音合成系统进行了多次评测,在评测中自然度评测采用的主要是听音人打分的方法。由于这种主观方法费时费力,而且受人主观的影响灵活性、稳定性和重复性都不高。为了弥补主观评测方法的这些缺点,我们研究了通过计算机对合成语音进行客观评测的方法作为主观评测方法的补充。通过前面所述的方法,我们通过 MATLAB 工具实现了对合成语音和自然语音的 MFCC 特征参数距离进行计算,从而获得了客观评测结果。将其与主观评测所得到的主观印象打分(MOS)的结果进行了对比分析,证明了通过分析合成语音和自然语音间的 MFCC 特征参数距离,所获得的客观评测结果与主观评测结果的相关度达到了 0.80 以上,由此证明通过 MATLAB 工具可以实现对合成语音自然度进行客观评测的目的。利用这个办法我们通过使用 MATLAB 工具实现了针对语音合成系统所输出语音的质量进行自动客观评价的评测系统。

参考文献

1 蔡莲红、黄德智、蔡锐等,现代语音技术基础与应用,清华大学出版社, 2003。

2 何强、何英,MATLAB 扩展编程, 清华大学出版社, 2002。

3 赵红怡、张常年,数字信号处理及其 MATLAB 实现,化学工业出版社,2002。