语音合成大作业

王炜致 2022010542

1 语音预测模型

(1)

给定

$$e(n) = s(n) - a_1 s(n-1) - a_2 s(n-2)$$

假设 e(n) 是输入信号，s(n) 是输出信号，上述滤波器的传递函数是什么？如果 $a_1 = 1.3789$，$a_2 = -0.9506$，上述合成模型的共振峰频率是多少？用 zplane ，freqz ，impz 分别绘出零极点图，频率响应和单位样值响应。用 filter 绘出单位样值响应，比较和 impz 的是否相同。

①上述滤波器的传递函数为

$$H(z) = \frac{1}{1 - \frac{a_1}{z} - \frac{a_2}{z^2}} = \frac{z^2}{z^2 - a_1 z - a_2}$$

②共振峰频率

$$f = \frac{\omega}{2\pi}$$

模拟频率 ω 和数字频率 Ω 有关系 $\Omega = \omega T$, 求解二次方程，得到

$$\Omega \approx \arctan \frac{0.68939}{0.68945} \approx 0.7854$$

故

$$f = \frac{\Omega}{2\pi T} = \frac{0.7854}{2\pi} \times 8000 \approx 1kHz$$

③用 zplane,freqz 分别绘出零极点图、频率响应如下（代码见附件，下同）：

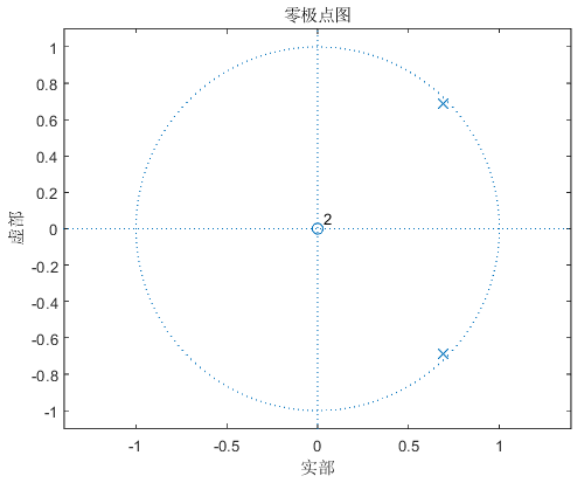

图 1: 零极点图

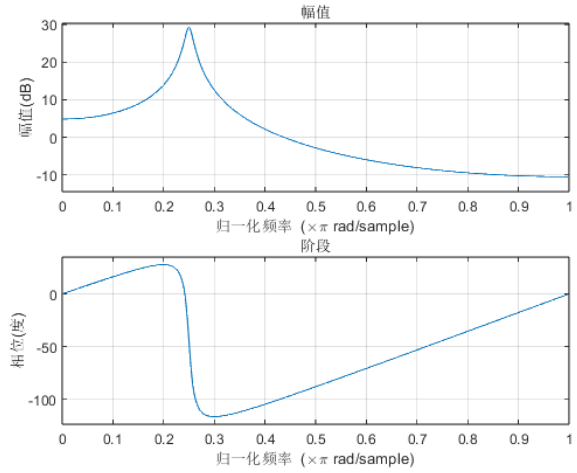

图 2: 频率响应

用 impz,filter 分别绘出单位样值响应如下，两者相同：

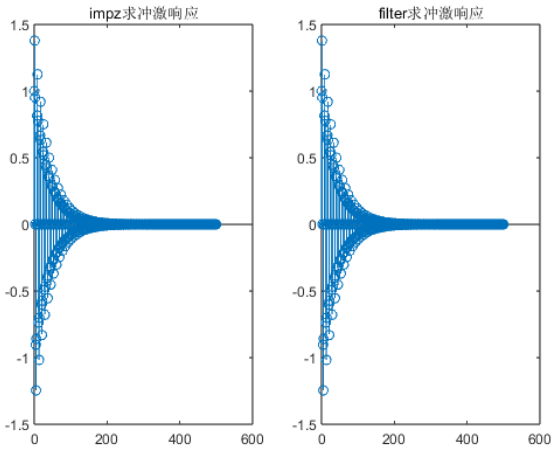

图 3: 冲激响应

(3)((2) 略)

运行程序到 27 帧时停住，用（1）中的方法观察零极点图。

如下所示，用 zplane 函数绘制 E,A 决定的零极点图即可。注意 A 是预测系数，应用以表示 z 变换表达式的分母部分。

```
% (3) 在此位置写程序，观察预测系统的零极点图
    zplane(E,A);
```

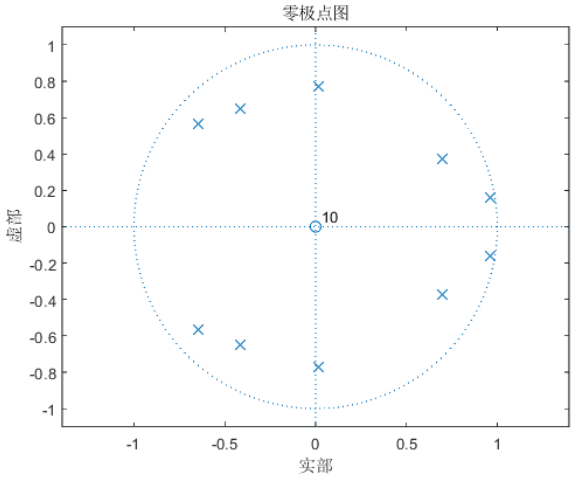

图 4: 27 帧对应零极点图

(4)

在循环中添加程序：对每帧语音信号 $\mathbf{s(n)}$ 和预测模型系数 $\{a_i\}$，用 **filter** 计算激励信号 $\mathbf{e(n)}$ 。

由于分帧处理，需要保存滤波器的最终条件 zf，且需要考虑滤波器的初始状态 zi，在代码中表现为更新 zi_pre 的值。注意求激励即预测残差需要将滤波器输入输出对换，A 应作分子，1 应作分母。

```
% (4) 在此位置写程序，用filter函数s_f计算激励，注意保持滤波器状态
    [exc_ftr,zi_pre] = filter(A,1,s_f,zi_pre);
    exc((n-1)*FL+1:n*FL) = exc_ftr;
% exc((n-1)*FL+1:n*FL) = ... 将你计算得到的激励写在这里
```

(5)

完善 speechproc.m 程序，在循环中添加程序：用你计算得到的激励信号 e(n) 和预测模型系数 $\{a_i\}$ ，用 filter 计算重建语音 $\hat{s}(n)$ 。

类似地有（注意输入输出调转）：

```
% (5) 在此位置写程序，用filter函数和exc重建语音，注意保持滤波器状态
    [rec_ftr,zi_rec] = filter(1,A,exc_ftr,zi_rec);
    s_rec((n-1)*FL+1:n*FL) = rec_ftr;
% s_rec((n-1)*FL+1:n*FL) = ... 将你计算得到的重建语音写在这里
```

(6)

在循环结束后添加程序：用 sound 试听（6）中的 e(n) 信号，比较和 s(n) 以及 $\hat{s}(n)$ 信号有何区别。对比画出三个信号，选择一小段，看看有何区别。

信号所载信息均为“电灯比油灯进步多了”。e(n) 信号噪声较大；s(n) 信号和 $\hat{s}(n)$ 信号听起来几无区别，且两者噪声相对 e(n) 较小。因为 e(n) 是语音信号 s(n) 和 $\sum_{k=1}^{N} a_k s(n-k)$ 的差（残差），自然与 s(n) 相差较大；但作为激励信号可以较好重建语音。

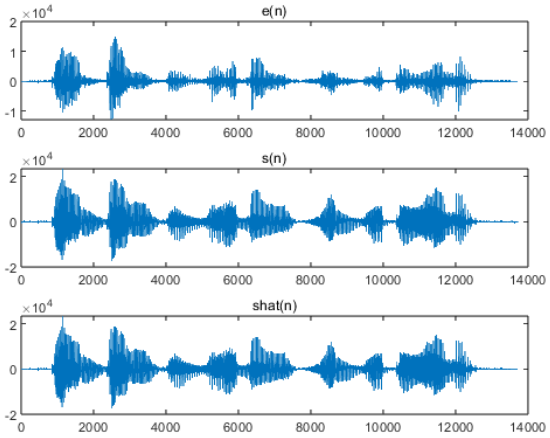

图 5: 语音信号比较

选取 [2000,4000] 局部区间比较，进一步验证了前述判断：

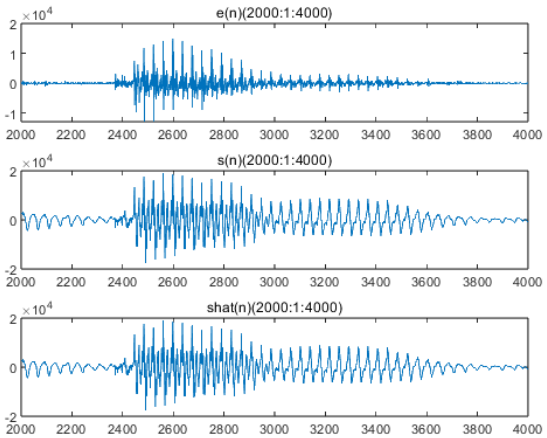

图 6: 语音信号比较（局部）

2 语音合成模型

(7)

生成一个 **8kHz** 抽样的持续 **1** 秒钟的数字信号，该信号是一个频率为 **200Hz** 的单位样值“串”，即

$$x(n)=\sum_{i=0}^{NS-1}\delta(n-iN)$$

考虑该信号的 **N** 和 **NS** 分别为何值？用 **sound** 试听这个声音信号。再生成一个 **300Hz** 的单位样值“串”并试听，有何区别？

①抽样频率的定义为单位时间内从连续信号中提取并组成离散信号的抽样个数。故 1s 内的抽样点有 8000 个；该信号频率为 200Hz，故 $NS=200, N=\frac{8000}{200}=40$。定义抽样频率

```
fsp=8000;
```

则用 sound 试听声音信号的代码为 (注意到 sound 默认声音时长为 1s)

```
seq1=zeros(fsp,1);
f1=200;
pulse1=floor(fsp/f1);
for n=1:f1 %matlab start from 1
    seq1(pulse1*n)=1;
end
sound(seq1);
```

②同上，信号频率为 300Hz 时，$NS=300, N=\frac{8000}{300}=26$, 试听代码为

```
seq2=zeros(fsp,1);
f2=300;
pulse2=floor(fsp/f2);
for n=1:f2 %matlab start from 1
    seq2(pulse2*n)=1;
end
sound(seq2);
```

注意到，300Hz 信号的音调显著地比 200Hz 音调要高。

(8)

真实语音信号的基音周期总是随着时间变化的。我们首先将信号分成若干个 **10** 毫秒长的段，假设每个段内基音周期固定不变，但段和段之间则不同，具体为

$$PT=80+5mod(m,50)$$

其中 **PT** 表示基音周期，**m** 表示段序号。生成 **1** 秒钟的上述信号并试听。

由于段长为 10ms，总长为 1s，易知信号的段数为 100，则信号生成并试听的代码为

```
fsp=8000; N=100;
seq=zeros(fsp,1); n=1;
while n<=fsp
    seq(n)=1;
    m=ceil(n/(fsp/N));
    PT=80+5*mod(m,50);
    n=n+PT;
end
sound(seq);
```

相邻两个脉冲间的 PT 值由更靠前的那个脉冲（所在的段号）决定。

(9)

用 filter 将（8）中的激励信号 e(n) 输入到（1）的系统中计算输出 s(n)，试听和 e(n) 有何区别。

试听比较，发现有显著变化，通过图像进一步观察其变化情况。将 e(n) 视为一系列冲激信号的叠加，则 s(n) 则对应为各冲激信号响应的叠加。各个冲激信号及其响应情况与 [图 3：冲激响应] 所示大体上是吻合的。

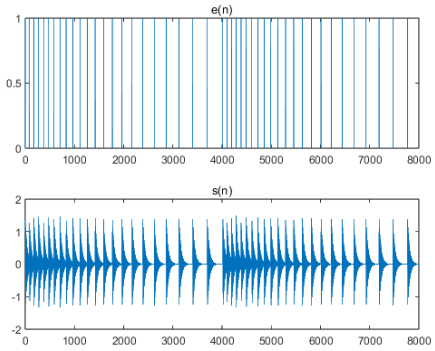

图 7: 输入 e(n) 与输出 s(n) 信号比较

(10)

重改 speechproc.m 程序。利用每一帧已经计算得到的基音周期和（8）的方法，生成合成激励信号 Gx(n)（G 是增益），用 filter 函数将 Gx(n) 送入合成滤波器得到合成语音 s (n) 。试听和原始语音有何差别。

帧长 FL=80。模仿 (8) 中代码，在循环开始前预置 k=2FL+1(即从第 3 帧开始)：

```
k = 2*FL + 1;
```

k 作为下标，为合成激励信号 exc_syn 在适当位置添加加权脉冲，具体操作如下：

```
while k <= FL*n
    exc_syn(k) = G;
    k = k + PT;
end
```

k 按照已经计算好的 PT 跳跃，在当前帧添加加权脉冲，当超过当前帧范围时结束，完成当前帧激励信号的合成。

将当前帧激励信号输入滤波器（注意保存状态），即得合成语音 $\tilde{s}(n)$：

```
[s_syn((n-1)*FL+1:n*FL),zi_syn] = filter(1,A,exc_syn((n-1)*FL+1:n*FL),zi_syn);
```

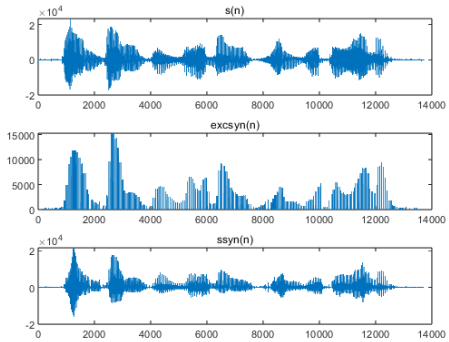

图 8: 合成信号比较

听觉上，s(n) 与 $\tilde{s}(n)$ 无明显不同，观察图像发现 s(n) 局部幅度更大些，其余差异亦不显著。

3 变速不变调

(11)

仿照（10）重改 speechproc.m 程序，只不过将（10）中合成激励的长度增加一倍，即原来 10 毫秒的一帧变成了 20 毫秒一帧，再用同样的方法合成出语音来。

将 n 统一改为 2n 即可:

```
% (11) 不改变基音周期和预测系数，将合成激励的长度增加一倍，再作为filter
% 的输入得到新的合成语音，听一听是不是速度变慢了，但音调没有变。
while k_v <= FL*n*2
    exc_syn_v(k_v) = G;
    k_v = k_v + PT;
end

[s_syn_v(2*(n-1)*FL+1:2*n*FL),zi_syn_v] = filter(1,A,exc_syn_v(2*(n-1)*FL+1:2*n*FL),zi_syn_v);
```

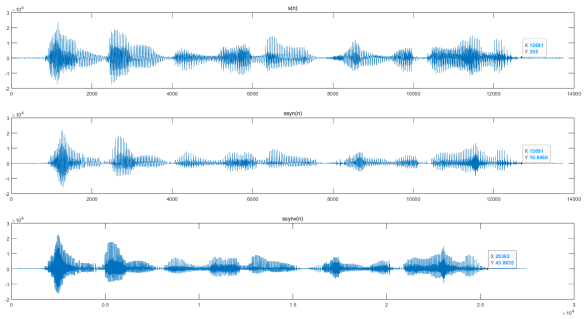

图 9: 变速信号比较

听取语音，记录两个语音 (s,s_v) 播放的大致时间，以及观察图像横坐标，可知信号长度确实增加了一倍; 此外音调并无明显变化，说明实现了变速不变调。

4 变调不变速

(12)

重新考察（1）中的系统，将其共振峰频率提高 150Hz 后的 a_1 和 a_2 分别是多少？

由于原系统共轭极点为 $z = 0.68945 \pm 0.68939j = 0.975e^{\pm j0.7854}$，而利用

$$f = \frac{\Omega}{2\pi T}$$

有

$$\Delta f = \frac{\Delta\Omega}{2\pi T}$$

给定 $\Delta f = 150Hz$，则

$$\Delta\Omega \approx 0.11781$$

对每一个极点乘上因子 $e^{\pm j\Delta\Omega}$ 即可按需改变共振峰频率。注意因子指数的正负与原指数正负一致。由此得到新系统共轭极点为

$$z = 0.975e^{\pm j0.9032} = 0.6036 \pm 0.7657j = \frac{a_1 \pm j\sqrt{-4a_2 - a_1^2}}{2}$$

解得 $a_1 = 1.2072, a_2 = -0.9506$。

(13)

重新考察（1）中的系统，将其共振峰频率提高 **150Hz** 后的 a_1 和 a_2 分别是多少？

首先将基音周期减半，只需引入新的 k_t 以及改为 $PT/2$:

```
while k_t <= FL*n
    exc_syn_t(k_t) = G;
    k_t = k_t + ceil(PT/2);
end
```

处理共振峰频率可以采用（12）中的方法。现需要改造滤波器。原有的滤波器由 [1,A] 表示，需要修改为 [B,T] 以改变共振峰频率。先后使用 tf2zp 和 zp2tf 函数可以完成操作。

```
[z,pole,r] = tf2zp(1,A);
pole_add = pole.*exp(1j*0.11781*sign(imag(pole)));
[B,T] = zp2tf(z,pole_add,r);

[s_syn_t((n-1)*FL+1:n*FL),zi_syn_t] = filter(B,T,exc_syn_t((n-1)*FL+1:n*FL),zi_syn_t);
```

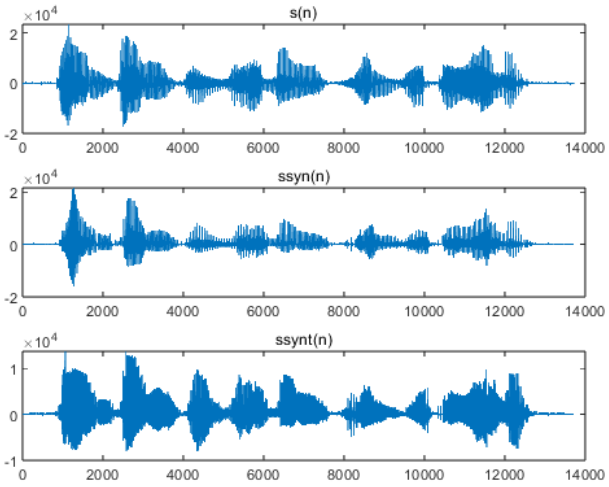

图 10: 变调信号比较

听取语音，发现明显由低沉男声变为尖细女声，即频率提高；观察图像，发现时长相等，即没有变速。这就实现了变调不变速。