音频转换实验 Part III

2014011355 辛杭高

2017 年 5 月 26 日

本部分实验要求对两段音频的距离进行测定，由于在实验 2 中，我主要使用了频率特征来修改从 A 到 B 的语音信息，因此在本实验中，我给出的距离度量准则主要是基于基频和频率的。

测试时运行 src 文件夹下的 test.m 文件即可，也可以直接在 doc/result.txt 中查看运行结果。注：本实验为了获取音频的基频信息使用了开源库 fxrapt，其源代码在 tools 文件夹下。

1 基频之差

我们可以使用基频之差的绝对值作为基频，用公式表述为

$$dis = |F_0(wav_1) - F_0(wav_2)| \tag{1}$$

具体代码参见 meandis.m，调用该函数之前应该首先调用 fxrapt 来获取相应音频的基频。直接使用基频之差作为距离在有些情况下有些武断，对于这种度量方式来说，基频是一段音频最重要的特征，对于实验 part2 中男女音频转化类的实验会比较有用。

在此距离度量下，郭德纲的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	86.6426
改变频率	9.1581
改变时长	12.2042

在此距离度量下，单田方的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	77.8518
改变频率	0.1616
改变时长	11.5972

在此距离下，A 类音频与 B 类音频的距离，A 类′ 音频与 B 类音频的距离度量后如下：

待转化文件名	A 与 B	A 与′B
sen6000	124.3862	22.8660
sen6015	113.6207	20.5444
sen6028	111.4070	33.0249
sen6044	102.4995	13.7764
sen6147	105.7581	29.6283

从实验 part1 的数据来看，仅仅改变基频的情况下，距离相比于改变频率和改变时长的变化更大。在 part1 中，我们了解到时长和频率之间有密切的关系，从这一度量数据下，也可以反映出这一结论。

在此距离度量下可以明显看出，A 与 A 的′ 距离明显小于 A 与 B 的距离，这主要是因为基频是本度量的唯一标准。因为从 A 转化到 A 后′ 距离变小，所以更加接近于音频 B。

2 基频概率分布之差

我们可以将原始的声音序列划分成多个子序列，并且统计各个子序列中基频出现的频率，再用频率的概率分布之差的欧式距离作为我们的度量准则。相对于直接使用基频度量来说，这种方式相当于将时域信息考虑了进来，所以从这个角度来说，这种距离度量更加客观一些。

用公式表述为

$$dis = \|P(F_0(wav_1)) - P(F_0(wav_2))\| \tag{2}$$

在此距离度量下，郭德纲的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	0.2709
改变频率	0.0390
改变时长	0.0398

在此距离度量下，单田方的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	0.1936
改变频率	0.0389
改变时长	0.0403

在此距离下，A 类音频与 B 类音频的距离，A 类´ 音频与 B 类音频的距离度量后如下：

待转化文件名	A 与 B	A 与´B
sen6000	0.5199	0.2281
sen6015	0.5047	0.2206
sen6028	0.4497	0.1914
sen6044	0.5202	0.1648
sen6147	0.4936	0.2692

同于前一种度量方法，对于所有文件 A 类与 B 类的距离要大于 A 类´ 与 B 类的距离，这可以解释为什么 A 更´ 接近于 B，但是从理论上来说，这种方法由于考虑到时长的因素，所以相较于第一种方式会更加科学一些。

3 fft 下的频域差距

这种度量的思路是，首先将音频从时域转化到频域，然后再频域上度量两者的距离，由于我们并不是很关心声音的强度，所以在具体实现的时候，我对 fft 转化后的结果首先进行了归一化，用公式可以表述为

$$dis = \|fft(wav_1) - fft(wav_2)\| \tag{3}$$

在此距离度量下，郭德纲的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	163.0106
改变频率	198.1963
改变时长	142.2059

在此距离度量下，单田方的音频与实验 part1 修改后的距离度量如下：

待转化文件名	距离
改变基频	83.3480
改变频率	154.2692
改变时长	185.0048

在此距离下，A 类音频与 B 类音频的距离，A 类′ 音频与 B 类音频的距离度量后如下：

待转化文件名	A 与 B	A 与′B
sen6000	48.3889	25.4606
sen6015	117.7738	102.8801
sen6028	71.4647	42.4653
sen6044	70.9075	54.6097
sen6147	77.4202	95.6550

首先对于郭德纲和单田方的音频，在本度量下改变频率成为造成距离增大的最主要原因，这是因为 fft 本身就是对频域的表征。从 A 类声音到 B 类声音的转换来说，对于绝大多数情况还是可以保证，A 与 B 的距离要大于 A 与′B 的距离，但对于 sen6147 来说，这个对比关系并不成立，将 sen6147 的频域图象绘制出以后发现，可以看出 A 的′ 频域分布整体上更接近与 B，但是出现了个别奇点，这些奇点导致最后度量的结果很大。

4 错误尝试–时间序列度量

还有一种比较直观的思路是直接在时域上对两段音频的距离进行度量，这种度量用公式可以表述为

$$dis = \|wav_1 - wav_2\| \tag{4}$$

显然在这种度量下，即使两段音频的频域信息十分类似，也可能在时域下距离很大。若将此度量准则用于 A 类语言与 B 类的比较，可以得到

待转化文件名	A 与 B	A 与′B
sen6000	0.0372	0.0387
sen6015	0.0125	0.0146
sen6028	0.0184	0.0206
sen6044	0.0154	0.0177
sen6147	0.0106	0.0131

此时，A 与 B 的距离反而小于 A 与′B

的距离，这违反了人的直观听觉，这说明这种方法在此数据集上并不合适，虽然这种方法在时域上也是有其道理的。

5 声学参数与人主观听觉的关系

对于人的主观听觉来说，声音的内容非常重要，如果两个音频的内容不同，那度量的距离结果应该很大。但声音的频率与内容的关系不大，所以在基频与频率的度量下很难度量出内容的不同，所以对于音频内容的识别还需要考虑到时域的信息。最好首先对音频进行划分，音频被分为多个片段，如果每个片段只包含少数文字，再对各片段声音的频域差异进行度量，就可以在一定程度上反映两个音频之间内容上的关联程度。当然，这种判断方法仍有局限，一方面是因为文字与声音的频率并不是一一对应关系，另一方面频域除了后所读文字的影响，还会受到说话人的影响。

从音色的角度来说，频域是度量音色的一个非常有效的指标，例如男女音色之间的差距，可以利用两者的频率进行度量，也可以通过频域上的调整将男音转化成女音。因此我们可以通过傅里叶变化，将声音从时域转化到频域，如果两个音波在频域下比较接近，那可以在一定程度上说明这两段音波的音色比较接近，尤其是针对男音和女音的辨识，有着很良好的效果。

基频是基音的频率，决定了整个音的音高，会影响人类主观听觉对自然中声音的感知，十分适合作为距离度量时的主要参数。

时长对应于音速特征，但是音速特征并不是独立的，改变时长会在一定程度上影响到频率，但改变时长依然是在改变音频的音色，而不是在改变音频的内容，只会使得音频变得更加低沉或者更加尖亮。

若想改变音频的内容，通过改变基频，频率，时长等特性并不可行，若要改变音频的内容，必须要在小片段上改变其声学特征，从这个角度来说，若是用小波变换来对音频进行处理，会同时考虑到频域特征和时域特征，会与人的主观听觉产生高度重合，可以作为一个重要指标。