基于 matlab 的变声器实现

信号 8 组
崔文彦 张讯 李昊霖 彭瑞南

一．概述：

在人们的日常生活中，变声器具有非常广泛的应用。电视台经常针对某些事件的知情者进行采访，为了保护知情者，经常改变说话人的声音。并且经常有情况需要将男性的声音变为女性的声音或女性的声音变为男性的声音。我们需要经过处理之后的语音信号并不影响收听和理解。我们所做的程序在满足上述条件的同时并且可以随意实现人声转换以及改变语速的快慢。

二．背景知识介绍

1.1 发音器官

人体的语音是由人体的发音器官在大脑的控制下做生理运动产生的。人体发音器官由三部分组成：肺和气管、喉、声道。 肺是语音产生的能源所在。气管连接着肺和喉，是肺与声道的联系通道。喉是由一个软骨和肌肉组成的复杂系统，其中包含着重要的发音器官——声带。声带为产生语音提供主要的激励源。声道是指声门（喉）至嘴唇的所有发音器官，包括咽喉、口腔和鼻腔。

1.2 语音的产生

语音是声音的一种，是由人的发声器官发出，具有一定语法和意义的声音。大脑对发音器官发出运动神经指令，控制发音器官各种肌肉运动从而振动空气从而形成。空气由肺进入喉部，经过声带激励，进入声道，最后通过嘴唇辐射形成语音。

2.1 基音周期和基音频率

基音周期的概念：人在发音时，声带振动产生浊音（清音由空气摩擦产生）。浊音的发音过程是：来自肺部的气流冲击声门，造成声门的一张一合，形成一系列准周期的气流脉冲，经过声道（含口腔、鼻腔）的谐振及唇齿辐射最终形成语音信号。故浊音波形呈现一定的准周期性。所谓基音周期，就是对这种准周期而言的。它反映了声门相邻两次开闭之间的时间间隔或开闭的频率。

基音周期是语音信号最重要的参数之一，它描述了语音激励源的一个重要特征。基音周期信息在语音识别、说话人识别、语音分析与语音合成，以及低码率语音编码、发音系统疾病诊断、听觉残障者的语言指导等多个领域有着广泛的应用。

3.1 语音信号的采样

采样率：频率对应于时间轴线，振幅对应于电平轴线。波是无限光滑的，弦线可以看成由无数点组成，由于存储空间是相对有限的，数字编码过程中，必须对弦线的点进行采样。采样的过程就是抽取某点的频率值，很显然，在一秒中内抽取的点越多，获取得频率信息更丰富，为了复原波形，一次振动中，必须有 2 个点的采样，人耳能够感觉到的最高频率为

20kHz，因此要满足人耳的听觉要求，则需要至少每秒进行 40k 次采样，用 40kHz 表达，这个 40kHz 就是采样率。常见的 CD，采样率为 44.1kHz。

采样大小：光有频率信息是不够的，我们还必须获得该频率的能量值并量化，用于表示信号强度。量化电平数为 2 的整数次幂，我们常见的 CD 位 16bit 的采样大小，即 2 的 16 次方。

采样率和采样大小的值越大，记录的波形更接近原始信号。

拓展：为什么采用 44.1kHz？

首先，44,100 = 2 ^ 2 × 3 ^ 2 × 5 ^ 2 × 7 ^ 2 因此，对于许多计算来说，44.1 kHz 实际上是一个很容易处理的数字。

其次，在 20 世纪 70 年代，当数字录音还处于初级阶段时，使用了许多不同的采样率，包括 Soundstream 录音中的 37 千赫和 50 千赫。在 70 年代后期，飞利浦和索尼合作研发了 cd，两家公司之间关于采样率有很多争论。最终，44.1 千赫被选中，原因有很多。根据奈奎斯特定理，44.1 千赫兹允许复制低于 22.05 千赫兹的所有频率内容。这包括一个正常人所能听到的所有频率。尽管对于高频内容的认知仍存在争议，但人们普遍认为很少有人能听到 20 千赫以上的音调。这张 44.1 千赫的 CD 格式的创作者还可以将至少 80 分钟的音乐(比黑胶唱片的时间还要长)装在一张 120 毫米的光盘上，这被认为是一个强大的卖点。

4.1PCM（Pulse Code Modulation)脉冲编码调制

脉冲编码调制是数字通信的编码方式之一。主要过程是将话音、图像等模拟信号每隔一定时间进行取样，使其离散化，同时将抽样值按分层单位四舍五入取整量化，同时将抽样值按一组二进制码来表示抽样脉冲的幅值。包括采样、量化和编码。

PCM 音频流的码率=采样率值×采样大小值×声道数 bps。采样率为 44.1KHz，采样大小为 16bit，双声道的 PCM 编码的 WAV 文件，它的数据速率则为 44.1K×16×2 =1411.2 Kbps。将码率除以 8，就可以得到这个 WAV 的数据速率，即 176.4KB/s。这也意味着一首 3 分钟的音乐将占用 30M 的数据空间，这是无法让人接受的（如今存储容量不断增加，越来越多的人为了追求绝对音质已经不在乎这个问题了，但在前些年，确实如此）。

4.2 有损、无损以及音频压缩

根据采样率和采样大小可以得知，相对自然界的信号，音频编码最多只能做到无限接近，至少目前的技术只能这样了，相对自然界的信号，任何数字音频编码方案都是有损的，因为无法完全还原。在计算机应用中，能够达到最高保真水平的就是 PCM 编码，被广泛用于素材保存及音乐欣赏，CD、DVD 以及我们常见的 WAV 文件中均有应用。因此，PCM 约定俗成了无损编码，因为 PCM 代表了数字音频中最佳的保真水准，并不意味着 PCM 就能够确保信号绝对保真，PCM 也只能做到最大程度的无限接近。我们而习惯性的把 MP3 列入有损音频编码范畴，是相对 PCM 编码的。

就像上文说的那样，无损音乐的空间占用很大，因此需要进行压缩。要降低磁盘占用，只有 2 种方法，降低采样指标或者压缩。降低指标是不可取的，因此专家们研发了各种压缩方案。由于用途和针对的目标市场不一样，各种音频压缩编码所达到的音质和压缩比都不一

样。有一点是可以肯定的，他们都压缩过。

下图是蓝牙通信中的通信协议、采样率已及量化位数，原理不同。

三．基频提取

下图为男性女性平均基频

	平均基频(Hz)	基频方差	最大基频(Hz)	最小基频(Hz)
男性平均值	160.81±24.27	1.23±0.38	165.73±23.21	159.40±22.17
女性平均值	297.42±35.89	2.42±0.74	306.26±37.87	293.30±38.11
总平均值	206.35±70.77	1.63±0.77	212.57±72.56	204.03±69.50

接下来我们考虑用自相关法提取信号的基频

原理：对于离散的语音信号 x(n)，它的自相关函数定义为：

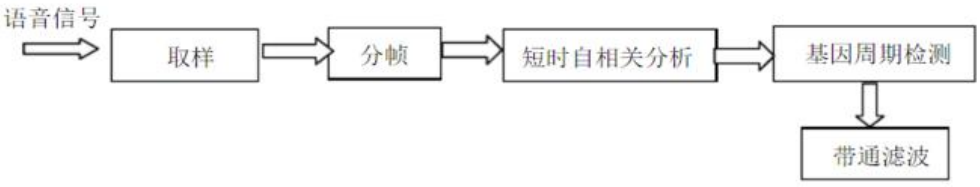
$$R(k)=\Sigma x(n)x(n-k),$$

如果信号 x(n))具有周期性，那么它的自相关函数也具有周期性，而且周期与信号 x(n)的周期性相同。自相关函数提供了一种获取周期信号周期的方法。在周期信号周期的整数倍上，它的自相关函数可以达到最大值，因此可以不考虑起始时间，而从自相关函数的第一个最大值的位置估计出信号的基音周期，这使自相关函数成为信号基音周期估计的一种工具。

而语音信号是非平稳的信号，所以对信号的处理都使用短时自相关函数。短时自相关函数是在信号的第 N 个样本点附近用短时窗截取一段信号，做自相关计算所得的结果

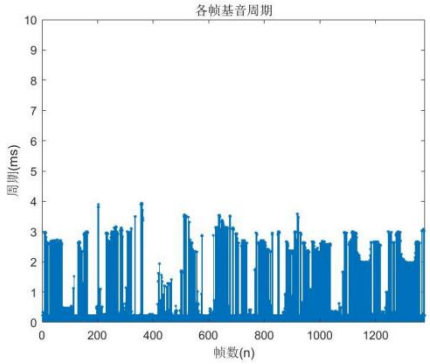
$$Rm(k)=\Sigma x(n)x(n-k)$$

该式中，n 表示窗函数是从第 n 点开始加入。

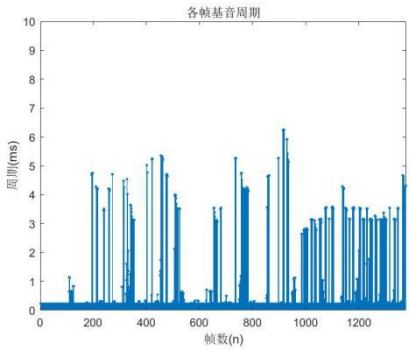


实验步骤：

1. 取一段录音作为音频样本
2. 对样本音频进行采样
3. 对采样后的音频进行分帧
4. 对每一帧求短时自相关函数
5. 算出对应周期

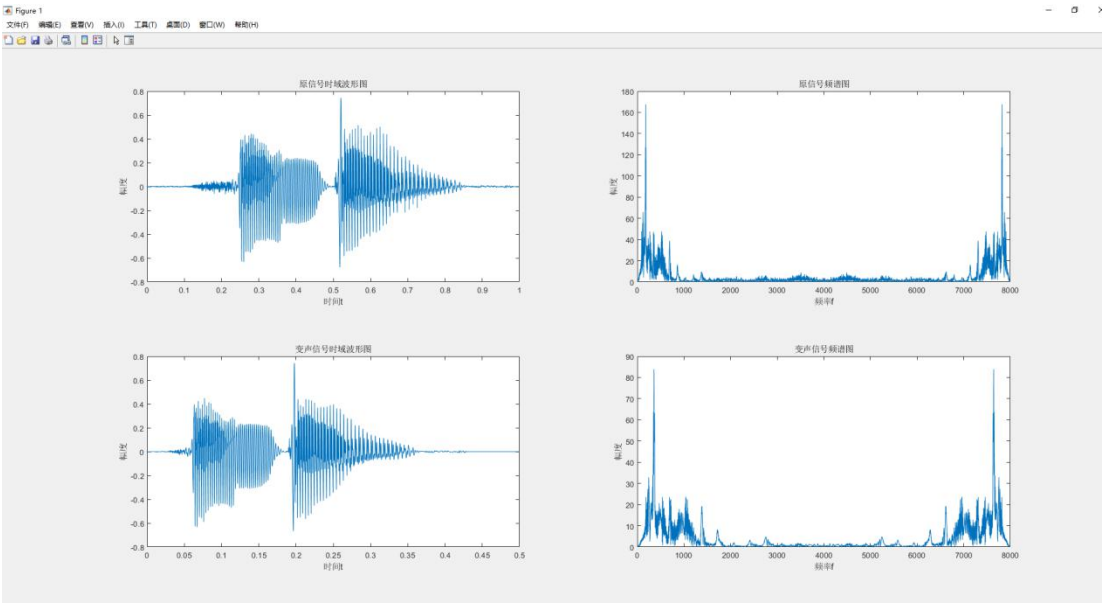


邓紫棋



陈奕迅

实验中将男声片段转化为了女声片段，效果下图所示



四．变声实现

1. 实现方法

变声器通过对音频信号进行重采样并规整时长得到变声后的信号。通过改变 voice 函数的参数 f(或改变参数 n)来改变采样率，f 大于 1 时为上采样，对应时域插值，频域压缩。而输入参数为 1 的时候，f=1 的时候为原声。f 小于 1 为下采样，对应时域抽取频域延拓，故偏向女声。

利用 matlab 自带的 resample 函数进行重采样，以实现声音信号频率的改变，但随之而来的是时长的改变，即信号的语速发生了变化。因此，需要通过重叠叠加算法来进行时长规整。对原始信号以固定的帧率 a 进行分解，再以固定帧率 b 进行合成，两者比值保持不变，进而保证了重叠区域的幅度一定。

代码，脚本详见附件。

2.例子

下述相关参数位置已在脚本中注释。

①男变女，女变男

首先对音频信号“爱情转移—陈奕迅”进行预处理，滤除背景音乐，只保留 30s 左右的人声，并利用脚本 zx2 提取得到陈奕迅的基频，以便于后续实现指定目标的变声。在脚本 zx 中改变 voice 函数的参数 n(f 的放大倍数，初始值为 1000，此时为原声)，使其小于 900，即可实现向女声的变换。

当小于 600 以后，声音较为尖锐的电音，使听者难以分辨男女，该频段的变声效果可以应用于保护说话人。

对音频信号“多远都要在一起—邓紫棋”进行上述的预处理。改变参数 n,使其大于 1200，即可得到男声，随着 n 值的增大，声音会越来越粗犷。

②语速改变可通过改变时长规整因子 F 实现，F 变大时语速加快，反之减慢。

③指定目标的变换

对于指定目标的声音变换效果较差，下面列举其中效果较好的两组。

陈奕迅—梁静茹：在提取了两者的基频后，根据比例调整参数，想要实现从陈奕迅到梁静茹的变换需要将 n 值设定为 840，实际测得 n 值为 830 左右效果更好（主观上），得到的音频有部分噪声。

同时，按比例改变参数 n 使其=1204，可将声音变回陈奕迅，但有噪声。

梁静茹—张信哲：过程如上述，n 值设定为 1100 左右，反变换 n 取 910 左右。

上述变声后的音频文件以 wav 格式保存在附件中。

五．总结

本次课程设计让我们受益匪浅，不仅对声音信号，时频域变换有了更深的了解，还学到了很多 matlab 的使用知识。我们的变声器项目依旧存在着很多不足，如变声后不可避免地产生噪声，对指定目标的变换效果较差等，这些问题值得后续进一步探究。