0 深圳技术大学考试答题纸

(以论文、报告等形式考核专用)

二〇 二四 ～二〇 二五 学年度第 二 学期

课程编号	JC00021	课程名称	数字信号处理	主讲教师	廖聪维	评分	
学 号	2021003030028	姓名	钱嘉涛	专业年级	微电子 1 班		

教师评语:

题目: 我的数字信号处理实验课成果

论文要求 3500 词以上，结构完整（包括标题、摘要、引言、结果与讨论、结论及参考文献），排版规范。

基于音高转变的语音性别转换算法

钱嘉涛 [1]

| （1. 深圳技术大学　集成电路与光电芯片学院，广东省　深圳市　518000；） |

摘　要：声音是人类常用的工具，是相互传递信息的最主要的手段，因此语音信号处理在现代信息社会中占有重要地位。本文主要讨论语音信号处理中的性别变换，其主要任务是对语音进行某种变换使之产生性别变化的特效，如在男声、女声、老年人声之间互相转换，达到伪装效果。本文介绍了一种基于环形缓冲区的音高转换算法，实现了高效的语音性别转换。该方法通过对输入音频信号进行分段处理，利用不同步长进行采样读取，来改变声音的音高，进而实现语音性别转换。

关键词：数字信号处理；语音性别转换；音高转换；环形缓冲区；

Speech Gender Conversion Algorithm Based on Pitch-Shifting

Jiatao Qian[1]

| （1. COLLEGE OF INTEGRATED CIRCUITS AND OPTOELECTRONIC CHIPS, Shenzhen Technology University, 518000, Guangdong Province, China；） |

Abstract： Sound is a common tool used by human beings, making it the primary means of information exchange. Therefore, it holds a significant position in modern digital signal processing. This paper mainly discusses gender transformation in speech signal processing. The main task is to apply certain transformations to the speech to produce gender-changing special effects, such as mutual conversion between male voice, female voice, elderly voice, to achieve disguise effects. This paper introduces a Pitch-Shifting algorithm based on Ring-Buffer, which achieves efficient speech gender conversion. This method changes the pitch of the sound by segmenting the input audio signal and sampling it at different step lengths, thereby achieving speech gender conversion.

Key words： Digital signal processing(DSP); Voice gender conversion; Pitch shifting; Ring Buffer;

1 引言

语音性别转换 (Voice Gender Conversion)，就是尽可能找出两个信号 (源说话人信号和目标转换信号) 声音特征之间的差别，在转换过程中改变源说话人的声音特征，使转换后的声音携带更多目标信号的声音特征，仿佛是原始目标信号发出的样，并且在转换过程中，源说话人的语义内容，环境信息等得以保留。

语音转换 [8]，是一个典型的多学科交叉的数字信号处理技术，其涉及信号处理、人工智能，模式识别，声学等学科领域，是继语音识别、说话人识别、说话人确认技术后语音信号处理领域新兴的又一研究热点，具有广泛的应用前景。例如在娱乐应用中，可以通过实时调高音高使得声音听起来尖，有 ‘‘小孩’’、‘‘外星人’’ 等听觉效果；

自 1970 年以来, 很多研究人员采用各种各样的转换技术以合成期望目标说话人的语音。Atal[1] 等人研究了使用 LPC 声码器改变声音特性的可行性。Childers[2] 等人检验了男女声互相转换的方法。Rinscheid[5] 使用时变滤波器和拓扑特征映射实现了声音的改变。Valbret[7] 等人使用基音同步叠加法调整激励信号中的韵律特征来改善转换性能。近年来，更多的研究人员致力于语音特征的统计分布来实现声音的转换，一些学者通过概率方法 [6]，采用高斯混合模型 (GMM) 描述源与目标特征的联合概率分布，这样给定源特征矢量寻找转换函数来预测目标语音特征就变成一个回归问题。GMM 技术比局部变换方法有效性、鲁棒性好，其原因在于对频谱包络建立了一个连续概率模型。GMM 联合概率方法理论上能使回归问题的混合成分得到更合理的配置，但在进行 EM 运算时计算量较大，而且存在频谱过分光滑的现象，影响了转换语音的目标倾向性。以上讲述到的方法虽然在语音细节与效果上具有一定优势，但是有一个共性的问题，即实现复杂，计算量大，难以实现转变的实时处理。由此小计算量的音高转换（Pitch Shift）可以提高语音性别转变的速度，从而实现快速的语音性别转变，甚至可以相对轻松地部署音高转换算法在硬件如 FPGA 开发板上。

音高转换是一种常用的实现语音性别转换的方法，声音的音高与声音的频率集合相对应。通过调整声音的频率成分。例如，通过调高一个男人唱歌的音高，我们可以得到一个听起来像是女人在唱歌的声音。

一个著名的音高转换的早期实践者是 Chuck Berry，他使用的技术，使他的声音听起

来更年轻。甲壳虫乐队在 1966 年和 1967 年的许多唱片都是通过录制器乐曲目制作的，这些曲目高半个音阶，而声音则相应地低一些。例如”Rain”，”I’m Only Sleeping” 还有”When I’m Sixty-Four”。电子音乐家 Burial 以在他的歌曲中包含声音旋律的音高变化样本而闻名。Goregrind 使用的声音，往往是基调转移到声音不自然的低沉和喉音。从 1986 年到 1988 年，美国音乐家普林斯使用音高转换创造了他的 “卡米尔” 人声。由此得出，基于音高转换而实现声音信号转变是很常见的。这一过程可以通过多种方法实现，包括但不限于采样率修改、谐波分析/重合成、相位编码器方法以及基于频谱的方法。我们通过使用两个环形缓冲区 (Ring-Buff) 来实现对音频的数据进行平滑以及不同步长的读取和写入。因为不同的读取步长，数据被下采样或重复采样，使得音频具有音高偏移效果。从而实现语音性别转换的功能。

2 方法

2.1 音高转变原理

2.1.1 乐理与音色:

在音乐中，每个八度包含 12 个半音，这是基于十二平均律的定义，即一个八度被等分为 12 个相等的半音。这种分割方式使得每个半音之间的频率比是 2 的 12 次方，即大约 1.05946。每个半音对应一个特定的音符。纯音符由基频的单个正弦波组成。图 1显示了最常见的音符及其相应的基频。

在生活中，不同乐器演奏的同一个音符不会产生相同的声音，这主要是因为每个乐器的谐波成分不同，从而导致音色差异。每个音符通常由一个基频和一组谐波组成。基频是指声音中最低的频率，而谐波则是基频的整数倍频率。每种乐器在演奏时，其基频虽然相同，但谐波成分却有所不同。谐波的频率和相位的不同，即物体振动分量的关系不同，决定了声音的音色。因此，如果我们想要把一个声音转换为另一个目标声音，可以比作为音色的转换，也就是频率的调整。

2.1.2 频率调整:

音高变化包括将旋律向上或向下移动一个或多个半音，如下式所示。初始半音指数为

Note	Hz	Note	Hz	Note	Hz	Note	Hz	Note	Hz	Note	Hz	Note	Hz
C1	32.7	C2	65.4	C3	130.8	C4	261.6	C5	523.3	C6	1046.5	C7	2093.0
C#1	34.6	C#2	69.3	C#3	138.6	C#4	277.2	C#5	554.4	C#6	1108.7	C#7	2217.5
D1	36.7	D2	73.4	D3	146.8	D4	293.7	D5	587.3	D6	1174.7	D7	2349.3
D#1	38.9	D#2	77.8	D#3	155.6	D#4	311.1	D#5	622.3	D#6	1244.5	D#7	2489.0
E1	41.2	E2	82.4	E3	164.8	E4	329.6	E5	659.3	E6	1318.5	E7	2637.0
F1	43.7	F2	87.3	F3	174.6	F4	349.2	F5	698.5	F6	1396.9	F7	2793.8
F#1	46.2	F#2	92.5	F#3	185.0	F#4	370.0	F#5	740.0	F#6	1480.0	F#7	2960.0
G1	49.0	G2	98.0	G3	196.0	G4	392.0	G5	784.0	G6	1568.0	G7	3136.0
G#1	51.9	G#2	103.8	G#3	207.7	G#4	415.3	G#5	830.6	G#6	1661.2	G#7	3322.4
A1	55.0	A2	110.0	A3	220.0	A4	440.0	A5	880.0	A6	1760.0	A7	3520.0
A#1	58.3	A#2	116.5	A#3	233.1	A#4	466.2	A#5	932.3	A#6	1864.7	A#7	3729.3
B1	61.7	B2	123.5	B3	246.9	B4	493.9	B5	987.8	B6	1975.5	B7	3951.1

图 1 音符及其基本频率

p_{inital}，用于移位的半音个数为 s，最终半音指数为 p_{final}。公式（1）为音高转换的初始半音和最终半音索引之间的关系。

$$p_{final} = p_{inital} + s \tag{1}$$

从信号的角度来看，这包括将基频和谐波按特定因子缩放。初始音符频率为 f_{inital}，移位的半音数为 s，最终音符频率为 f_{final}。公式（2）为音高转换的初始频率和最终频率之间的关系。

$$f_{final} = 2^{(s/12)} \times f_{inital} \tag{2}$$

举个例子，当我们将音符向上或向下移动一个完整的八度时，实际上是在频谱上进行了缩放操作。前面提到过每个半音之间的频率比是 2 的 12 次方，如果我们将音符向上移动一个八度，那么每个音符的频率都会翻倍，因为从一个八度的 C 到另一个八度的 C，频率增加了 2 倍。同样地，如果我们将音符向下移动一个八度，每个音符的频率都会减半，因为从一个八度的 C 到另一个八度的 C，频率减少了一半，即频谱缩放。

2.2 算法实现原理

2.2.1 错误原理:

录制一段音频，然后让它以两倍的速度播放，那么所有的频率都会加倍，音高也会上移一个八度。从图2可看出信号被缩短了一倍。

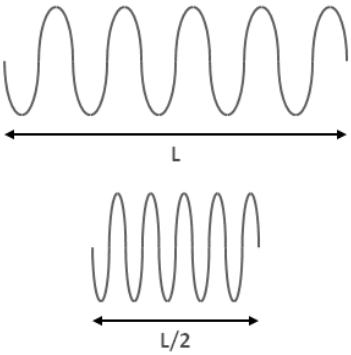

图 2 通过影响持续时间来改变音高

2.2.2 基准原理:

把这段音频在不影响音高的情况下将长度加倍，然后以两倍的速度播放。此时。所有频率都将加倍，因此音高将发生变化，并且持续时间将与初始频率相匹配。图3为音频处理过程简图。

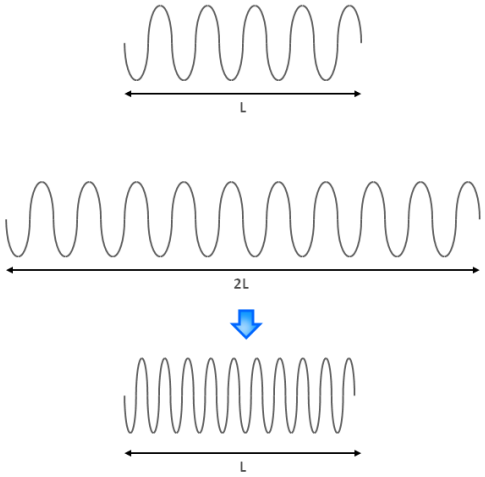

图 3 在不影响持续时间的情况下进行音高转换

2.3 程序设计

程序通过 MATLAB2020b 设计仿真。考虑到方便程序未来在硬件上的移植，我们采用手动分配缓冲区地址的方式以及双缓冲区的设计，这样的设计不仅能够自由改变音频的音高，还能创造出独特的声音效果，如和声以及平滑过渡。

2.3.1 环形缓冲区:

环形缓冲区是一种固定大小的数据结构，它在音频处理中扮演着至关重要的角色。在本算法中，我们定义了四个环形缓冲区，每个缓冲区的大小为 B = 1024。这些缓冲区用于存储音频信号的样本，并允许我们以循环的方式进行读写操作。通过模运算 mod 来实现索引的循环，确保当索引超出缓冲区大小时能够自动回绕到起始位置。

2.3.2 音高转换:

音高转换是通过改变音频信号的频率来实现的。在本算法中，我们定义了两个音高转换的增量因子 delta1 和 delta2，它们分别用于调整不同缓冲区中音频样本的读取位置。这些增量因子是通过计算 2 的幂来得到的，其中幂的底数为 -3/12 和 -6/12，这对应于音乐理论中的半音音程。通过这种方式，我们可以在不改变原始音频长度的情况下，实现音高的升高或降低。

关于 delta 值的设置，可以通过修改底数 2 或修改幂值甚至同时修改来设置音高转换的方向以及程度。其底数决定了基础音高的倍数。例如，如果将 2 改为 0.5，则表示基础音高是缩小为原来的 1/4。修改底数会对音高产生较大的影响，通常不会通过修改这个系数来微调音高，因为这会导致音高变化过大。修改幂值能够实现音高的精细调整。例如，修改 -3/12 为 -2/12，则表示音高会升高一个半音（1 个半音 = 1/12 个倍频程）。通过修改幂值，可以更加精细地控制音高的升降，这对于声音的变换来说是非常重要的，使用这种方法，可以实现半音级别的精确调整。

2.3.3 缓冲区的读写:

对于每个输入样本，我们将其写入四个缓冲区中。写入操作通过计算模索引来完成，确

保数据在缓冲区中循环存储。同时，我们根据音高转换的增量因子更新读取索引，并通过模运算 mod 计算出实际的读取地址。

2.3.4 音频信号的合成:

在读取了不同缓冲区中的样本后，我们通过加权平均的方式合成新的音频信号。加权合成能够创造出和声效果。这种合成方法不仅能够保持音频的连续性，还能够起到增加音频的丰富性和表现力以及平滑音频的作用。

2.3.5 数据后处理:

在音频合成完成后，我们使用巴特沃斯低通滤波器并设置通带截止频率为 3500Hz 以去除高频噪声。这种滤波器能够有效地减少信号中的高频噪声，同时保持音频信号的人声部分，从而提高音频的质量。

2.4 实际设计中需要考虑的问题

2.4.1 噪声伪影与双缓冲区的设计意义:

当 delta>1 时，读取环形缓冲区的速率会高于写入速率，这意味着读取操作可能会超出当前写入的数据范围，从而读取到以前写入的“旧”数据。这会导致音频信号的不连续性，从而产生“咔哒”声和其他噪声伪影。为了缓解这一问题，通过设计双缓冲区，将第二个环形缓冲区的写入索引偏移总缓冲区大小的一半，这样可以使两个缓冲区的内容有 180 度的相位偏移。其效果类似拥有两个指针读取不同位置。

这样的设计能够平滑的音高变化。两个读取位置，它们分别对应原始音频信号中的不同相位。当一个指针从缓冲区的末尾“回绕”到开始时，另一个指针位于缓冲区的中间位置，这样可以在两个相位之间平滑过渡，从而实现连续的音高变化。

避免音频的不连续性。如果只有一个指针，当它到达缓冲区的末尾并回绕到开始时，可能会出现音频的不连续性，这会导致听觉上的“跳跃”或“点击”声。两个指针相隔 180° 的设置可以避免这种不连续性，因为当一个指针回绕时，另一个指针仍在缓冲区的另一端继续读取，从而实现平滑的过渡。

实现线性插值。通过两个指针，可以在它们之间进行线性插值，以生成新的样本点。这种插值方法可以改善音高变换后的音频质量，减少可能的失真和混叠效应。

交叉淡入淡出（Crossfading）。两个指针不仅用于读取和插值，还用于实现交叉淡入淡出效果。当一个指针在缓冲区的末端回绕时，通过逐渐减少该指针的权重并增加另一个指针的权重，可以创建一个平滑的过渡，减少听觉上的突兀感。

利用音频这类锯齿波形信号的周期性。锯齿波形被用来控制指针在缓冲区中的位置。由于锯齿波形是周期性的，当一个指针完成一个周期并开始新的周期时，另一个指针正好处于其周期的中间位置，这种设置利用了锯齿波形的这一特性来保持音高变换的连续性。

实现复杂音频效果。两个指针的设置还可以用于实现更复杂的音频效果，如和声、回声等，通过调整两个指针的相对位置和速度，可以创造出丰富的音频变换效果。

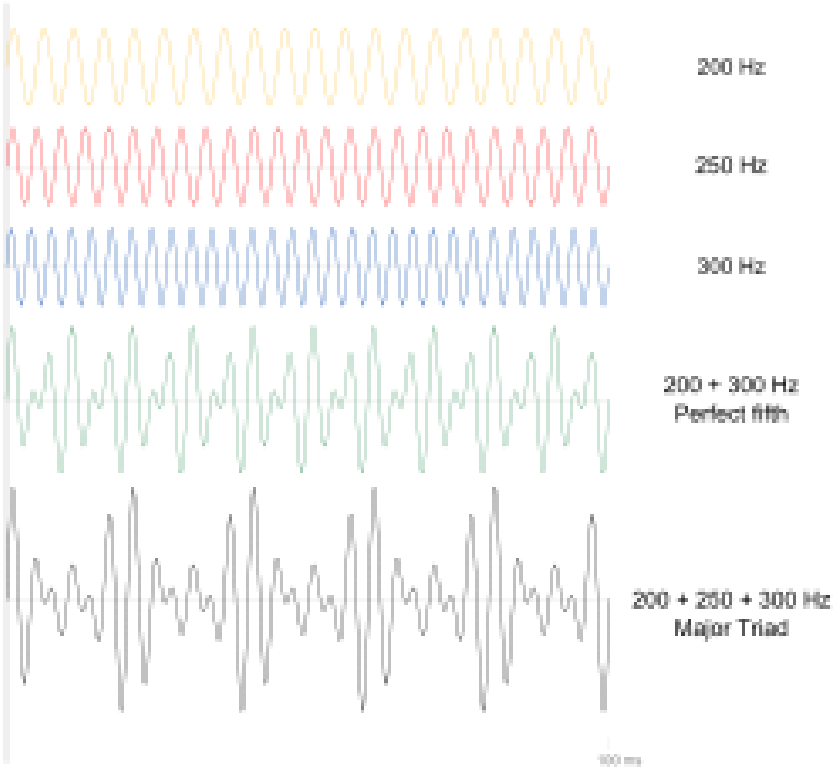

图 4 不同频率信号的和声频率示意图

3 结果

设置环形缓冲区的位数为 1024，为了得到显著的音高变化效果，增量因子的底数均为 0.25，幂值分别为-3/12 和-6/12，表示在相同幂值下，音高提高 3 个八度。通过 MATLAB 仿真，将得到如图5中的输入输出的时域图、频谱图还有时谱图。从六幅图中，我们能够在时谱图中较直观的看到处理后的频率更加平滑连续，其中高频噪声也由于低通滤波器而被滤除。

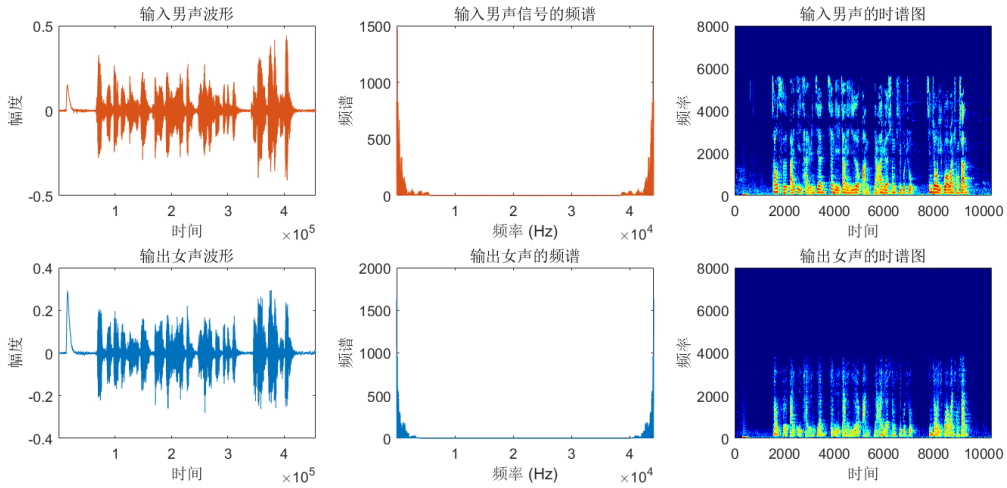

图 5 MATLAB 仿真结果

3.1 特征相似性评价指标计算：

我们分别通过时域波形欧氏距离、梅尔频率倒谱系数（Mel-frequency cepstral coefficients，MFCC）还有 Mel 频谱对比计算变换前后的相似性。计算同时要保证对比的数据长度一致且没有零值。通过公式（3）和公式（4）实现。

$$minLength = \min(\text{length}(s_{in}), \text{length}(filtered_s)) \tag{3}$$

$$signal = signal + 1e^{-20} \tag{4}$$

3.1.1 时域欧氏距离

欧氏距离是一种直观且易于理解的距离度量方法 [9]，它通过计算两个向量之间的平方

和来衡量它们的相似度。在音频处理中，时域波形图可以直接用作向量，因此使用欧氏距离进行比较非常直观和简单。计算方法如公式（11）所示。从结果上看，经过音高变换后的信号与原始信号在时域上具有一定程度的改变。

$$euclideanDistance = \sqrt{\sum_{i=1}^{minLength} (audio1[i] - audio2[i])^2} \tag{5}$$

3.1.2 MFCC 欧氏距离

欧氏距离不仅适用于时域特征，还可以用于频域特征以及时频特征的比较。这使得它在音频相似性度量中具有广泛的应用范围。例如，在语音识别中，常用的 MFCC 特征也可以通过欧氏距离进行比较。在 MATLAB 中 MFCC 的处理为一个封装好的函数，这里简单讲述一下其原理。主要步骤为: 预加重、分帧与加窗、通过傅里叶变换得到每帧总能量、通过 Mel 频率与 Mel 滤波器处理信号。

其中分帧的原因是，如果对整段语音做 FFT，就会损失时序信息。因此，我们假设在很短的一段时间 t 内的频率信息不变，对长度为 t 的帧做傅里叶变换，就能得到对语音数据的频域和时域信息的适当表达。为了帧与帧之间的连贯性每一帧的前 N 个采样点数据与前一帧的后 N 个采样点数据一样。图（6）为原理示意图。

Mel 频率的概念，Mel 值更加接近于人耳的听觉机制，其在低频范围内增长速度很快，但在高频范围内，Mel 值的增长速度很慢。每一个频率值都对应着一个 Mel 值，其对应关系如下公式（6）与示意图图（7），其中 m 为 Mel 值也叫 Mel 频率。

$$m = 2595 \cdot log_{10}(1 + \frac{f}{700}) \tag{6}$$

在确定 Mel 刻度和 Mel 滤波器个数后，谱线索引号 k 计算公式（7）如下，其中其中 N 为 FFT 点数，f_s 为抽样频率，f_m 为 Mel 刻度转化为频率后的值。

$$k = \frac{(1 + N) \cdot f_m}{f_s} \tag{7}$$

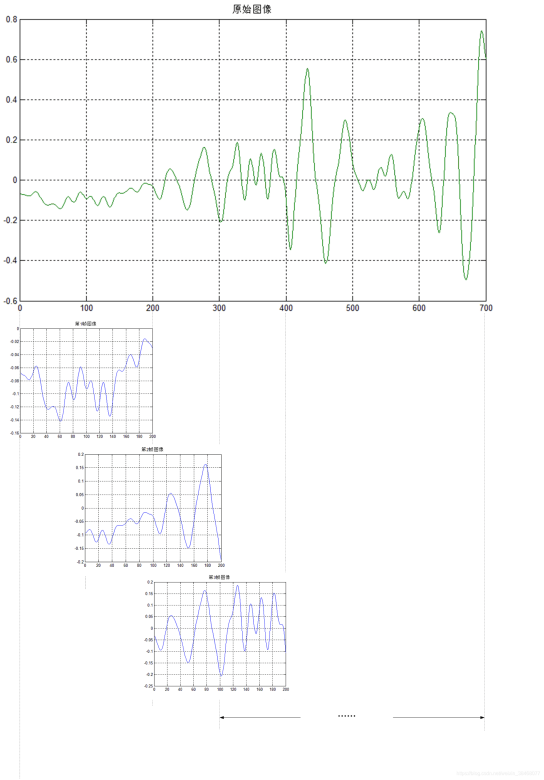

图 6 分帧示意图

最后再通过公式（8）求解 Hm 矩阵, 并进行离散余弦变换公式（9）

$$H_m(k) = \begin{cases} 0 & \mathrm{k} < f(m-1) \\ \frac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & f(m) \le k \le f(m+1) \\ 0 & k \ge f(m+1) \end{cases} \tag{8}$$

$$H = E \cdot H_m^T \tag{9}$$

就可以得到 MFCC 的计算公式（10）了。

$$mfcc(i,n) = \sum_{m=1}^{M} log[H(i,m)] \cdot cos[\frac{\pi \cdot n \cdot (2m-1)}{2M}] \tag{10}$$

在音频处理中，MFCC 特征能够捕捉人类耳朵对音频信号的感知特性 [4]，因此使用欧氏距离可以有效地评估两个音频文件之间的相似性。同样的从计算的结果来看，对于音频

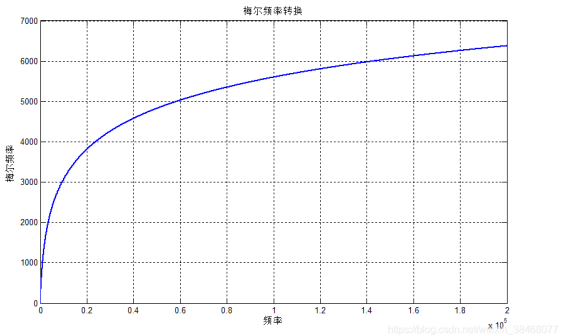

图 7 频率转换关系

中人类敏感的音频特征成分也具有一定的改变。

$$\text{mfccDistance} = \sqrt{\sum_{i=1}^{n} (\text{mfcc1}_i - \text{mfcc2}_i)^2} \tag{11}$$

3.1.3 Mel 谱相似性对比

Mel 谱中，可以看到的欧氏距离的差别是非常小的。Mel 谱展示了信号的能量在时间和 Mel 尺度上的分布，而 MFCC 则是对这些能量进行倒谱处理后得到的特征向量。这样结果说明了，转变后的信号在多尺度上的能量分布较一致，但是在 MFCC 中提取的特征向量又有一定差异，符合转变的特点。各相似性评价得分如表（1）所示。

表 1 变换后与原始数据的相似性得分

评价方法:	时域欧氏距离	MFCC	Mel 频谱对比
Score	47.04	185.39	0.0007385

3.2 噪声评价指标计算:

本部分对双缓冲区和巴特沃斯低通滤波器的降噪能力进行验证。通过对只用一个缓冲区以及用双缓冲区还有双缓冲区配合巴特沃斯低通滤波器这三个方式输出的音频进行信噪比计算结果如表（2）所示。由于没有纯净信号，我们分别通过自适应滤波器和基于直方图的噪声估计算法对噪声进行估计。

3.2.1 自适应滤波器计算信噪比

自适应滤波器估计噪声的实现为: 先定义输出 y 它是权重向量 w 和输入向量 x 的点积:

$$y(i) = w^T x \tag{12}$$

误差 e，它是输入信号 audio(i) 和输出 y(i) 的差:

$$e(i) = audio(i) - y(i) \tag{13}$$

以及更新权重 w，使用步长 u 、输入向量 x 和误差 e(i):

$$w = w + \mu x e(i) \tag{14}$$

估计噪声: 噪声 noise 是输入信号 audio 和误差信号 e 的差:

$$noise = audio - e \tag{15}$$

估计噪声功率: 噪声功率 noisePower 是噪声 noise 的平方的均值:

$$noisePower = mean(noise^2) \tag{16}$$

估计信号功率: 信号功率 signalPower 是输入信号 audio 的平方的均值:

$$signalPower = mean(audio^2) \tag{17}$$

计算 SNR，信噪比 snrValue 是信号功率 signalPower 和噪声功率 noisePower 的比值的 10 倍对数:

$$snrValue = 10\log_{10}\left(\frac{signalPower}{noisePower}\right) \tag{18}$$

表 2 LMS 计算信噪比

输出方式:	单缓冲区	双缓冲区	双缓冲区加低通滤波器
SNR(dB)	27.442	17.2335	20.1803

3.2.2 基于直方图计算信噪比

基于直方图估计噪声的实现为:

计算直方图: 假设音频信号为 x，通过直方图统计音频信号中各个强度级别的出现频率。我们将直方图的横轴分 N 份，即 N 个 bin。每个 bin 的范围即为信号大小范围 a_i, a_{i+1}，其中 i 是 bin 的索引。计算每个 bin 中的样本数，即得到直方图。

在进行噪声估计时，直方图的 bin 数目（也就是直方图中柱状的数量）对于音频信号的分析有很大的影响。这是因为 bin 的数量决定了算法对信号的分辨率，也就是区分信号中不同强度级别的能力。如果 bin 的数量过少，可能无法准确地估计信号中的噪声水平。这是因为过少的 bin 会导致多个不同强度级别的信号被归入同一个 bin，从而使得噪声水平的估计变得不准确。如果 bin 的数量过多，虽然我们可以得到更高的分辨率，但是也可能会引入更多的随机误差。这是因为过多的 bin 可能会导致一些 bin 中只包含了很少量的信号，从而使得噪声水平的估计受到这些随机误差的影响。因此，最终选择 110 作为直方图中 bin 的数量。

估计噪声水平: 我们假设噪声水平在直方图中最频繁出现的值附近。我们找到计数最大的 bin 的索引 i_{max}，然后计算这个 bin 的中心值作为噪声水平的估计，即 noise level。

$$\text{noise power} = (\text{noise level})^2 \tag{19}$$

噪声功率即噪声水平的平方。

$$\text{noise level} = \frac{a_{i_{max}} + a_{i_{max}+1}}{2} \tag{20}$$

计算信号功率: 信号功率可以通过计算音频信号的均方值得到，即 $signalpower$。

$$\text{signal power} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{21}$$

计算信噪比: 信噪比是信号功率和噪声功率的比值，通常以分贝 (dB) 为单位。

$$\text{SNR} = 10\log_{10}\left(\frac{\text{signal power}}{\text{noise power}}\right) \tag{22}$$

表 3 基于直方图计算信噪比

输出方式:	单缓冲区	双缓冲区	双缓冲区加低通滤波器
SNR(dB)	25.03	26.16	38.36

3.2.3 评价分析:

第一种方法通过自适应滤波器计算信噪比。从表格 (2) 中我们可以发现虽然双缓冲区在加上低通滤波器后确实信噪比有 3dB 的提高，但反常的是计算出来的单缓冲区的信噪比比双缓冲区的要高。这些是与我们预期相悖的结果。

第二种方法通过直方图计算信噪比。从表格 (3) 中我们发现双缓冲区相对单缓冲区信噪比有 1dB 程度的提高，因为其设计的本意是为了去除因为读取超程导致的类似 “哒哒” 的高频噪声伪影。符合我们的预期结果。当采用双缓冲区加低通滤波器时，信噪比更是有相对单缓冲区 13.3dB 的提高，因为直接滤除了 3500Hz 以上的不属于人声的高频噪声，其中过渡带到 4500Hz。也符合我们的预期结果。

根据分析，造成第一种方法结果异常的原因可能为:

第一，如果输入信号的噪声不是平稳的，直接从初始无语段来计算噪声，并假设后续噪声是稳态的进行谱减法，这种方法会导致误差。尽管在中间的无语段有所调整，但这种微调可能不足以解决问题;

第二，在对噪声观测值进行加权时，如果加权方式不恰当，会使滤波器估计值偏差较大，从而影响信噪比的准确性;

第三，如果输入信号存在相关性，前一次迭代产生的梯度噪声会传播到下一次迭代，造成误差的反复传播，收敛速度变慢，跟踪性能变差;

第四，模型和统计特性的不匹配：状态方程描述的动力学模型不准确，或者噪声的统计模型不准确，都会导致模型和量测值不匹配，从而影响信噪比的计算;

第二种方法选择基于直方图的噪声估计算法。其优势包括以下点:

鲁棒性：直方图方法对噪声的鲁棒性较强，因为它主要依赖于信号的能量分布，而不是信号的具体形状。这种方法可以有效地处理各种类型的噪声。

灵活性：直方图方法可以灵活应用于不同频带的噪声估计，通过分析各频带最常出现的能量值，可以准确反映每个频带的噪声水平。

无需先验知识：与自适应滤波器相比，直方图方法不需要关于信号统计特性的先验知识。自适应滤波器需要根据信号的实际情况不断调整滤波器参数，这在实际应用中可能会遇到困难。

简单性和直观性：基于直方图的噪声估计方法通过分析信号的能量分布来估计噪声水平，这种方法直观且易于实现。它不需要复杂的数学模型或大量的参数调整，适用于各种不同的噪声环境。

实时性：直方图方法通常计算速度较快，特别是在处理大规模数据时，能够快速生成噪声估计结果。这对于实时音频处理应用非常重要，可以减少延迟，提高系统的响应速度。

4 讨论

传统的音高转换方法需要复杂的数学模型和大量的计算资源。例如，Kang[3] 等人提出的基于 GMM 的方法需要通过复杂的统计模型来实现转换。这种方法在实际应用中可能会面临计算效率低下的问题。而本文通过改变频率使得改变音高的同时计算量较小。但是音高转换方法在保持语音内容信息不变的同时，可能无法完全捕捉到目标说话者的个性化特征，如基频、声道特征和韵律特征等。这导致转换后的语音在个性化表现上仍然显得生硬或不自然。所以音高转换技术仍然需要进一步优化以更好地处理多维度特征。

5 结论

本文通过设计双缓冲区并通过不同的读取步长实现快速的音高转换，同时对算法进行仿真与效果测试，成功实现了平滑的可调整的语音性别转换。

参考文献

[1] B. S. Atal and Suzanne L. Hanauer. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 08 1971.

[2] D. Childers, Ke Wu, and D. Hicks. Factors in voice quality: Acoustic features related to gender. In *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 293–296, 1987.

[3] Yongguo Kang, Jianhua Tao, and Bo Xu. Applying pitch target model to convert f0 contour for expressive mandarin speech synthesis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I, 2006.

[4] A. Maazouzi, N. Aqili, A. Aamoud, M. Raji, and A. Hammouch. Mfcc and similarity measurements for speaker identification systems. In *2017 International Conference on Electrical and Information Technologies (ICEIT)*, pages 1–4, 2017.

[5] A. Rinscheid. Voice conversion based on topological feature maps and time-variant filtering. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1445–1448 vol.3, 1996.

[6] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.

[7] H. Valbret, E. Moulines, and J.P. Tubach. Voice transformation using psola technique. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 145–148 vol.1, 1992.

[8] 符敏. 声音转换算法研究. 2006.

[9] 郭兴吉 and 范秉琪. 基于特征的音频比对技术. *河南师范大学学报 (自然科学版)*, (02):35–38, 2006.