电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

《数字信号处理》
课程设计论文

COURSE PROJECT THESIS

论文题目 DoGest：基于双频的超声波手势识别系统

学　　院 英才实验学院

专　　业 通信工程

学　　号 2017000203008 2017020901008 2017020913019

作者姓名 刘涵章 张天祺 蔡畅

指导教师 武畅

摘　要

近年来，凌空手势识别正成为一种与终端设备交互的流行方式。现在主要流行的终端设备有智能手机，平板电脑，笔记本电脑等，为了在这些终端上实现凌空手势识别，人们探索了各种利用不同的传播介质和设备的实现方案，包括基于计算机视觉的方案，基于 RE(射频信号)的方案，基于 WiFi(无线网络)的方案以及基于超声波的方案。在这些实现方案中，基于超声波的方案相比其他方案具有低成本，高鲁棒性等优点，因此正逐渐成为当今的研究热点。

在当前的超声波手势识别方案中，大多数方案利用一个声音频率，由此导致一些手势无法识别，比如左右滑动手势。针对这个问题，在本文中，我们主要构建了基于双频的超声波手势识别系统 DoGest。这个系统的实现硬件只需要借助笔记本电脑和智能手机里配有的扬声器和麦克风。主要的实现环境是超声波环境。其中 DoGest 利用笔记本电脑中的扬声器分别发出两个不同的高频声音信号实现了推，拉，左右滑动和转一圈、转三圈六个基本的手势。我们在室内环境中对该系统进行测试，，实验结果表明 DoGest 的平均手势识别准确率约为 94.5%。我们的主要贡献如下：

大多数基于超声波的方法只利用了一个声波频率，由此不可避免的限制了手势的种类，比如说难以识别出水平滑动手势，与此不同，DoGest 通过发出两个不同的声波频率信号(18kHz 和 19kHz)从而可以较快速的识别出左滑手势和右滑手势。而且相比于基于计算机视觉的方法，DoGest 可以不受光照条件影响的工作，不需要其它的特制硬件设备，且因为计算复杂度低具有低功耗的特点。

关键词：多普勒效应；凌空手势识别；超声波；手势分类

ABSTRACT

In—air gesture recognition is becoming a popular means of interacting with terminal devices，such as smart phones，tablets and laptops．To implement in-air gesture recognition，many methods have been researched，such as computer vision based，RF based，WiFi based and ultrasonic based．Compared to other approaches，the ultrasonic based method shows superiorities in low-cost, robustness and so on.

At present, most of studies utilize only single sound frequency, which limit the type of gestures, for example, they cannot distinguish swipe left and swipe right gestures. To solve this problem, in this dissertation, we construct DoGest, a dual-frequency based ultrasonic gesture recognition system. We mainly use the devices which are already embedded in laptops and phones including two speakers and one microphone．Our systems work on the ultrasonic environment．DoGest generates two different high frequency tones from two speakers and recognizes six gestures including push, pull, swipe left, swipe right, single circle and triple circle．For performance analysis, we test DoGest indoors, experimental results show that DoGest achieves about 94.5% average gesture recognition accuracy．Our contributions are as follows：

Most of ultrasonic based methods exploit single tone, which limit the type of gestures．For instance, they can not recognize horizontal swipe gestures．On the contrary, DoGest can recognize swipe left and swipe right gestures by using two high tones(18kHz & 19kHz)．Also compared to computer vision based method, DoGest can work under any light conditions without extra hardware devices and with low power consumption．

Keywords: Doppler Effect; In-Air Gesture Recognition; Ultrasonic; Gesture Classification

目 录

第一章 绪论

1.1 研究背景及现状

在人机交互领域的发展趋势中，以“人”为中心的人机交互技术势必要逐渐取代以“计算机”为中心的交互方式。在各种人机交互方式中，手势交互是最自然的方式。为了实现凌空手势识别，许多实现方案都已经被提出来过，根据不同的实现方案可以分为：基于计算机视觉的方案，基于数据手套的方案，基于惯性传感器的方案，基于深度传感器的方案，基于电磁波传感器的方案，基于超声波传感器的方案等等。基于计算机视觉的方案一直是实现凌空手势识别的主流方式。但是，基于视觉的方案仍然有许多限制和不足之处，比如，对光照条件敏感，太强或者太弱的光照条件都会降低设备识别手势的准确率，而且，还需要额外的深度传感器，此外，由于图像识别涉及的计算量常常会非常大，因此功耗也比较高。相比于这些方案，基于超声波的手势识别方案能够弥补上述的不足之处，它既不会受光照条件影响，也不会受普通噪音的影响，而且由于不需要图像识别，它的计算复杂度相对来说比较低，因此它的功耗相比基于计算机视觉的方案来说也更低。

1.2 研究内容及贡献

在本文中，我们提出并实现了基于超声波的手势识别系统 DoGest。这个系统依赖的硬件主要有笔记本电脑和智能手机中内置的扬声器和麦克风。在 DoGest 系统中，我们通过超声波信号的频谱变化来确定手势移动所引起的反射信号的多普勒频移方向，并根据不同的手势运动所产生的多普勒频移时间序列也不相同这一特点，将其作为手势特征，通过 k 近邻分类器，完成对手势的识别，DoGest 能够识别推，拉，左滑，右滑，转一圈，转三圈六个手势。DoGest 系统相比之前的系统，主要的改进之处在于：DoGest 可以识别出左滑和右滑手势，而之前的工作大多只采用了一个声波频率，正如 Fu et al.[1]指出，理论上，左右滑动手势单用一个声音频率是无法识别的，因为它们都是先靠近设备再远离设备，所以产生的频移序列是相似的(先正移，后负移)。Dolphin[2]只是借助左右滑动和单击手势所产生的能量分布不同，来作为区分这些手势的依据。但是，如果用户习惯使用左手进行操作，那么识别结果将是错误的，再者，每个用户的滑动习惯不一样，因此基于能量的方法在一些情况下并不一定有效。与此不同，我们利用笔记本内置的左

右两个扬声器分别发出两个不一样的超声波信号，当用户操作手势时，手势移动将会在这两个声波信号上同时产生频移，我们将同一时刻产生的两个多普勒频移组合成一个元组，比如说(1,-1)(1 表示手势靠近设备，-1 表示手势远离设备，元组的第一个元素和第二个元素分别代表左扬声器和右扬声器探测到的多普勒频移)。左滑手势将产生形如$[(1,1)^{+},(1,0)^{+},(1,-1)^{+},(0,-1)^{+},(-1,-1)^{+}]$的序列，右滑手势产生形如$[(1,1)^{+},(0,1)^{+},(-1,1)^{+},(-1,0)^{+},(-1,-1)^{+}]$的序列，这两个序列显然不一样，因此我们能够区分它们。我们把一个手势操作过程中所产生的元组的时间序列作为手势特征，并将其作为 k 近邻(kNN)分类器的输入，结合动态时间规整方法(DTW)，然后 kNN 输出手势识别的结果。实验结果表明 DoGest 对六种手势的平均手势识别准确率约为 94.5%。Ai et a1[3]利用手势相对设备的移动速度，手势频移能量大小，手势持续时间等作为手势特征输入至分类器来识别手势，但是，利用频移计算出的速度并不是手势运动的真实速度(除非手势运动轨迹正好处于手与设备的连线上)，该速度同时取决于手势运动速度和手与设备的相对角度，而且，手势持续时间与手势并无明显关系，仅取决于用户使用习惯，所以这些都不适合作为手势特征来考虑。这也是本文 DoGest 系统没有考虑将速度作为手势特征的原因。

第二章 基础知识

本文研究内容需要用到物理，数字信号处理，机器学习等领域的相关知识，为了便于理解后续的研究内容，因此，本章将依次介绍相关的知识理论。其中物理相关知识为多普勒效应；数字信号处理相关知识包括欠采样，滤波与窗函数，短时傅里叶变换；机器学习相关知识有动态时间规整，k 近邻分类。其中通过多普勒效应来检测手势移动的方向，利用欠采样提高信号的频谱分辨率，滤波和窗函数主要用于减少环境噪音的影响和频谱泄漏，动态时间规整用来衡量时序信号之间的相似度，k 近邻分类器用来对手势进行分类。

2.1 多普勒效应

在经典物理学中，多普勒效应是指当波源和接收者之间的相对距离发生变化且波源和接收者的移动速度相对于传播介质的速度比波在传播介质中的速度要低的时候，波源的频率f_0和接收者的接收频率f关系由如下公式给出：

$$f = \frac{c \pm v_r}{c \pm v_t} f_0 \tag{1}$$

其中

c为波在介质中的传播速度

v_r为接收者相对于传播介质的速度，如果接收者靠近波源，则为正值，如果远离波源则为负值

v_t为波源相对于传播介质的速度，如果波源远离接收者，则为正值，如果靠近波源则为负值

在本文中，扬声器作为发射器(声源)和麦克风作为接收器都是固定在笔记本电脑内部的，本身不会发生运动，唯一运动的物体只有手掌的移动。此时，可以把手掌近似的看作一个平面，该平面将入射的声波信号反射出去，于是移动的手势可以看作为移动的平面，持续的反射信号，从另一个角度看，对于入射的声波信号，该平面可以看作接收者接收入射的信号，对于反射的声波信号，该平面可以看作发送者发出反射的信号，因此该手掌平面作为接收者的同时也是发送者，假设手掌在反射入射的声波信号的时候，不会产生能量损耗。再假定手掌的移动速度为v，手接收到的入射信号的频率为f_r，在扬声器发出声波信号到手掌接收入射信号的过程中，只有接收者手掌在移动，发送者扬声器固定不动，将声音在空气中的传播速度记为c，可以得到手掌接收入射信号的声波频率为

$$f_r = \frac{c + v}{c} f_0 \tag{2}$$

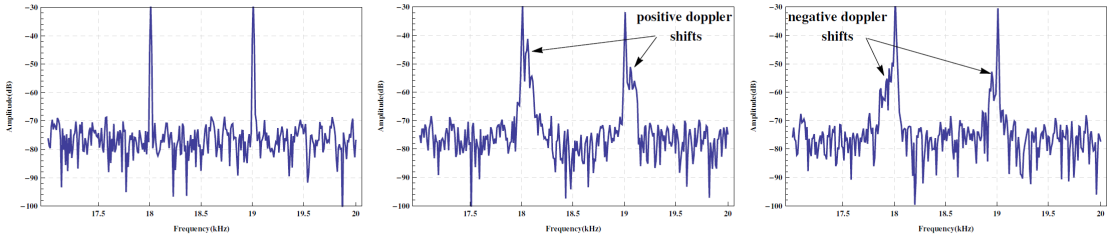
在手掌反射声波信号到麦克风接收声波信号的过程中，只有手掌作为发送者在移动，接收者麦克风固定不动，由此得到麦克风端接收到的声波频率为：

$$f_1 = \frac{c}{c-v} f_r = \frac{c+v}{c-v} f_0 \tag{3}$$

在实际的场景中，手掌的移动速度远小于声音在空气中的传播速度，即$c \gg v$，因此有

$$\Delta f = f_1 - f_0 = \frac{c+v}{c-v} f_0 - f_0 \approx \frac{2v}{c} f_0 \tag{4}$$

上式表明声音信号的频率变化与手掌的移动速度近似成线性关系，当手势运动为靠近设备，$\Delta f > 0$，反之，当手势运动为远离设备，$\Delta f < 0$。



(a) Stationary (b) Approaching (c) Departing

图 1 多普勒频移

因为在麦克风处不仅会接收经过手掌到达的声波信号，还会接收直接从扬声器发送过来的声波信号以及背景环境反射的声波信号，麦克风最终接收的信号是以上三部分声波信号的叠加。其中我们称从扬声器直接到达麦克风的传输路径为视线距离传播(1ine of sight)。图 1 展示了手掌移动时，麦克风接收到的声波信号频率分布，其中横坐标为频率，纵坐标为幅值，因为视线距离传播的能量损耗最小，所以其在麦克风的接收信号的频谱中对应着幅度最大的频率成分，此频率的大小和扬声器发出频率大小相同。从图 1(a)中可以看到，当手掌靠近麦克风时，麦克风接收到的声波信号存在大于原始声波信号频率的成分，图 1(b)中显示了远离手势使得接收到的信号有小于原始信号频率的成分。这些额外的频率成分即为多普勒频移。

2.2 欠采样

频谱分辨率Δn与 FFT 点数N和采样率f_s的关系如下：

$$\Delta n = \frac{F_s}{N_{FFT}} \tag{5}$$

为了提高频谱分辨率(使Δn更小)，有 2 种方式：降低采样率、提高 FFT 点数。由于增加 FFT 点数会带来更高的计算负担，我们采样欠采样技术降低采样率，提

高频谱分辨率。

麦克风的基带采样率$f_{SB}=48\text{kHz}$，选取的两个超声波信号频率为18kHz和19kHz。由于多普勒频移的大小在$\pm 200\text{kHz}$以内，我们分析超声波信号频率附近的带通信号即可，无需分析$[0\text{Hz}, 24\text{kHz}]$频率范围的基带信号。欠采样倍率$k$，采样率$f_s$与带通信号的频率范围$[f_L, f_H]$的关系如下：

$$\frac{2f_H}{k} \le f_s \le \frac{2f_L}{k-1} \tag{6}$$

表格 1 列出了欠采样倍率、采样率与通带范围和频谱分辨率的关系。我们选取欠采样倍率$k=6$.

表格 1 欠采样参数选取

k	$f_s=f_{SB}/k$	(f_L, f_H)	N_{FFT}	Δn
1	48K	(0, 24K)	4096	11.7
5	9.6K	(14.4K, 19.2K)	4096	2.3
6	**8K**	**(16K, 20K)**	**4096**	**2**
7	6.9K	(17.2K, 20.7K)	4096	1.7
8	6K	(18K, 21K)	4096	1.5

如表格 1 所示，6 倍欠采样时通带范围为$[16\text{k}, 20\text{k}]$. 抗混叠滤波器采用 IIR 设计，采取 18 阶 Butterworth 逼近，基本参数如下：

$F_{pass1}=18\text{kHz}$, $F_{pass2}=19\text{kHz}$, $F_{stop1}=16\text{kHz}$, $F_{stop2}=20\text{kHz}$, $A_{pass}=0\text{dB}$, $A_{stop}=-80\text{dB}$

该滤波器的幅频响应如图 2 所示。

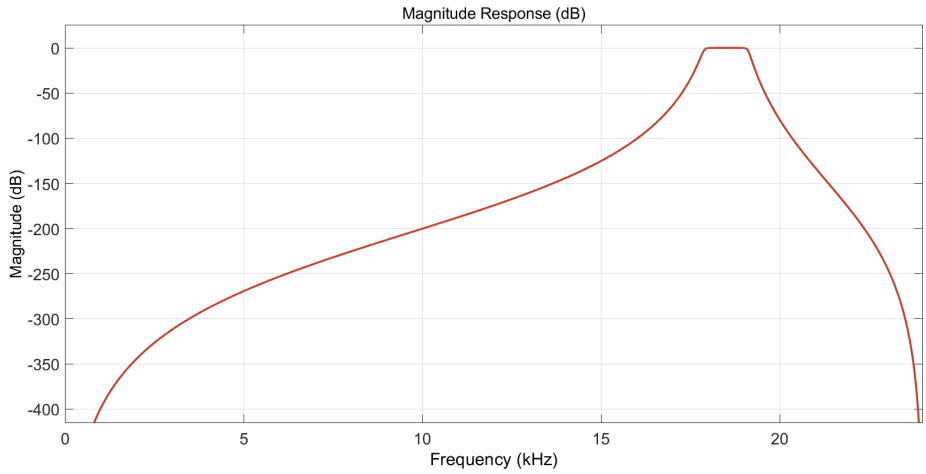


图 2 抗混叠滤波器的幅频响应

2.3 滤波与窗函数

当我们对原始信号采样后，我们得到的是一系列的振幅点的时间序列$s[n]$，理想情况下，$s[n]$是一个序列为无穷长的时间序列，在进行傅里叶变换的时候，我们不可能对这个无限长的序列做傅里叶变换，因为计算机只能同时处理有限个数据点。然而，矛盾的是，由于我们之前讨论的傅里叶变换都是在完整信号$s(t)$上进行的，这意味着，为了能够让计算机处理，不可避免地对$s[n]$的时间序列做截断。截断意味着原始信号失真。截断操作相当于时域上的离散信号$s[n]$乘上一个窗函数$w[n]$，其中窗函数是数据值个数为N的点集，形如$\{w[n], 0 \le n \le N一1\}$. 最理想的窗函数是不产生任何的频谱“泄漏”。但是只有频谱为$\delta[\omega]$的窗函数才不会改变原始信号的频谱，其对应的窗函数的时域表达式为$\{w[n], -\infty \le n \le +\infty\}$，为长度无穷长的常数序列，这在现实中这是不可能的。实际中一般选频谱尽可能接近于$\delta[\omega]$的窗函数。

考虑到时域乘积相当于频域卷积，所以我们需要窗函数具有特点：主瓣窄，旁瓣小。考虑一个最简单的截取，即直接对原始数据$s[n]$以每N个点作为分割，即乘以矩形窗函数。矩形窗的旁瓣电平起伏大，和原频谱卷积会产生较大的失真，习惯地我们也称之为频谱“泄漏”，因此矩形窗函数不是一个好的窗函数，在本文中，基于以上考虑，我们选取了汉明窗函数，其定义如下：

$$w[n] = (1-\alpha) - \alpha \cdot \cos\frac{2\pi n}{N-1}, 0 \le n \le N-1, \alpha = \frac{21}{46} \tag{7}$$

2.4 短时傅里叶变换

短时傅里叶变换的定义如下：

$$STFT\left(e^{j\omega}, n\right) = \sum_{m=-\infty}^{\infty} x[n-m]w[m]e^{-j\omega m}$$

其中，$w[n]$即 2.3 节所述的窗函数。

$$w_{Hamming}[n] = 0.54 + 0.46\,cos\left(\frac{2\pi n}{N_w}\right), -\frac{N_w}{2} < n \le \frac{N_w}{2} \tag{8}$$

2.5 动态时间规整

动态时间规整(DTW)是一种在时间序列分析种常用的度量两个时间序列相似度的算法。该算法最早 1968 年提出[5].在语音信号处理方面，经常用于补偿不同讲话人在语速上的区别。当不同讲话人发声同一个单词时，其在时间尺度上不一样，有人语速快，发音持续时间短，有人语速慢，发音持续时间长。但是不同人发出

的声音却有一定的相似性，如果使用诸如欧氏距离或夹角余弦的相似度判据，则测量的相似度是基于相同时刻信号的差异，不能抹除掉不关心的语速问题。

本文的手势识别也有相似特点，对于不同用户，就算他们执行相同手势，有些用户手势动作慢，持续时间长；有些用户手势动作快，持续时间短。本文中将使用 DTW 技术

DTW 是一个典型的优化问题，它用满足一定条件的的时间规整函数描述测试模板和参考模板的时间对应关系，求解两模板匹配时累计距离最小所对应的规整函数。若从当前的参考模板得到其特征矢量的序列，我们记成$\mathrm{A}=(i_1,i_2,\cdots,i_n)$，而从输入模板得到的特征矢量序列我们记成$\mathrm{B}=(j_1,j_2,\cdots,j_m)$. 对于每一对时刻$(i_s,j_s)$寻找一条最短的映射关系$P_0=(p_1,p_2,\cdots,p_k)$使得参考模板与输入模板之间在最优状态时的距离值$\mathrm{D(A,B)}$最短。

$$\mathrm{P}_0=\arg\min_{p_s} D(A,B)=\arg\min_{p_s}\frac{\sum_{s=1}^{k} d(p_s)\cdot w_s}{\sum_{s=1}^{k} w_s} \tag{9}$$

其中$d(p_s)$为点i_s,j_s之间距离，$w_s\geq 0$为权重因子。

规整函数p_s有如下性质

1. 单调性：单调性保证路径不会在时间轴上向之前的时刻行进。确保特征不会在路线中重复。表示为

$$i_{s-1}\leq i_s, j_{s-1}\leq j_s \tag{10}$$

2. 连续性：在时间轴上不会跳跃。确保规整方式不会忽略重要特征。表示为

$$i_s-i_{s-1}\leq 1, j_s-j_{s-1}\leq 1 \tag{11}$$

3. 边界条件：规整路径从左下角开始，到右上角结束。确保规整函数考虑了序列的全部特征。表示为

$$i_1=1, i_k=n, j_1=1, j_k=m \tag{12}$$

4. 窗口约束：良好的规整路径不太可能偏离对角线。确保规整路线不会尝试跳过不同的特征并卡在相似的特征上。表示为

$$|i_s-j_s|\leq r \tag{13}$$

其中$r>0$是窗口长度

5. 坡度约束：良好的规整路径不太可能过于平缓或陡峭。防止时间序列非常短的部分与非常长的部分匹配。表示为

$$\frac{j_{s_p}-j_{s_0}}{i_{s_p}-i_{s_0}}\leq \mathrm{p}, \frac{i_{s_q}-i_{s_0}}{j_{s_q}-j_{s_0}}\leq q \tag{14}$$

选用对称形式的权重因子

$$w_s = (i_s - i_{s-1}) + (j_S - j_{s-1}) \tag{15}$$

将上述公式表示为最优化的形式

$$\mathrm{D} = \frac{1}{\mathrm{m}+\mathrm{n}} \min_{p_s} \sum_{s=1}^{k} d(p_s) \cdot w_s$$

$$\text{s.t.} \begin{cases} i_{s-1} \le i_s \le i_{s-1} + 1 \\ j_{s-1} \le j_s \le j_{s-1} + 1 \\ i_1 = 1 \\ i_k = n \\ j_1 = 1 \\ j_k = m \\ |i_s - j_s| \le r \\ \dfrac{j_{s_p} - j_{s_0}}{i_{s_p} - i_{s_0}} \le \mathrm{p} \\ \dfrac{i_{s_q} - i_{s_0}}{j_{s_q} - j_{s_0}} \le q \end{cases} \tag{16}$$

若使用动态规划算法求解

$$g(i,j) = \begin{cases} d(i,j) + g(i,j-1), & i = 1, j > 1 \\ d(i,j) + g(i-1,j), & i > 1, j = 1 \\ 2d(i,j) + \min\{g(i-1,j), g(i,j-1), g(i-1,j-1)\} & i > 1, j > 1 \end{cases} \tag{17}$$

得到$g(n,m)$为 A、B 序列相似度。时间复杂度O(mn)

2.6 k 近邻分类

k 近邻(k-Nearest Neighbor，简称 kNN)分类是一种常见的监督学习方法，其工作机制非常简单：给定测试样本，基于某种距离度量找出训练集中与其最靠近的 k 个训练样本，然后基于这 k 个“邻居”的信息来进行预测，通常，在分类任务中可使用“投票法”，即选择这 k 个样本中出现最多的类别标记作为预测结果；在回归任务中可使用“平均法”，即将这 k 个样本的实值输出标记的平均值作为预测结果；还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大。在本文中，因为待识别的手势为分类任务，所以采用“投票法”。

第三章 DoGest 系统的实现

3.1 系统结构

本文中的 DoGest 系统主要包含三个子模块，包括数据采集和预处理模块，手势特征提取模块以及手势分类模块。在数据采集和预处理模块中，笔记本电脑通过内置的扬声器发送两个不同频率的声音信号，并且通过麦克风持续的录音采集信号，通过抗混叠滤波器并进行短时傅里叶变换，此模块主要完成声音信号数据的采样和预处理，使之成为手势特征提取模块可以处理的数据；在手势特征提取模块中，DoGest 通过次峰搜索算法计算出手势移动引起的多普勒频移，并提取出两个频率对应的频移作为二元组，将二元组的时间序列作为手势特征，完成手势特征的提取；在手势分类模块中，手势分类器先通过带标签的数据对分类器模型进行训练。然后将手势特征提取模块中提取出的手势特征作为该分类器的输入值，并得到输出值作为手势识别的结果。为提高系统的展示性与互动性，我们用推箱子的游戏与听众现场互动。图 3 展示了 DoGest 系统的总体结构。

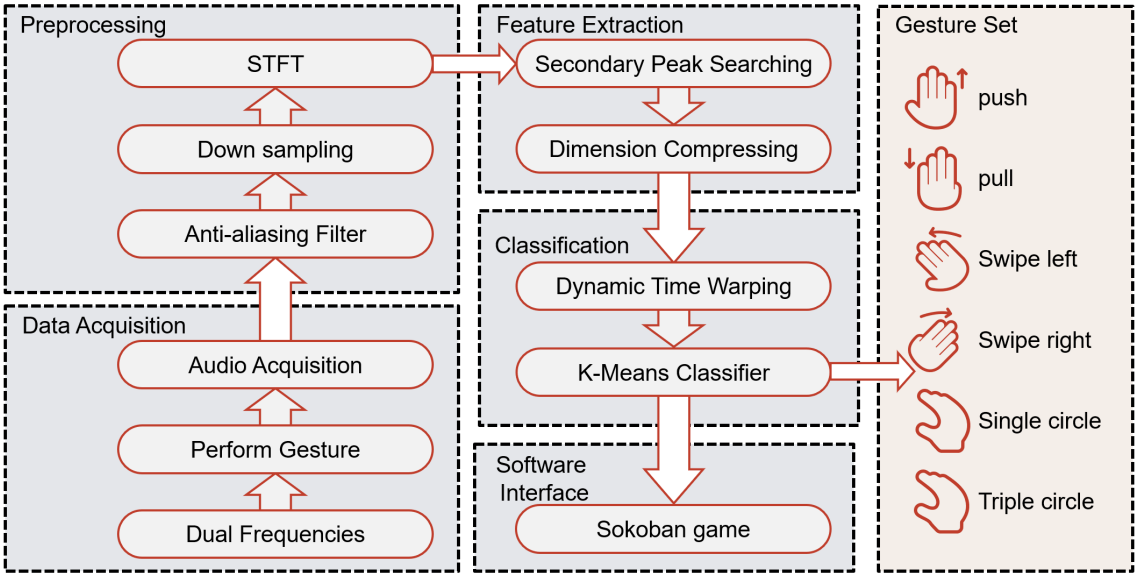


图 3 DoGest 系统的总体结构

从图中可以清楚的看出数据在不同模块之间的流向，如果将整个 DoGest 系统抽象为一个黑盒，那么这个黑盒的输入就是用户的手势操作，输出则是手势识别结果。

3.2 数据采集和预处理

在以往的实验中，大部分的研究只利用一个扬声器和一个麦克风来实现手势

识别，导致系统对手势移动信息获取有限，由此不可避免的限制了手势的种类，在本文中，我们利用笔记本电脑内置的两个扬声器和一个麦克风分别作为声波信号的发射器和接收器，额外多出来的一个扬声器将会提供手势移动更多的信息。

一般而言，人类可以听见的声音频率范围为 20Hz 到 20kHz，频率超过 20kHz 的声波我们称之为超声波。实际中，大部分人听不见频率大于等于 l8kHz 的声波，因此，在本实验中，我们选用的两个声波频率分别为 18kHz 和 19kHz。我们让位于笔记本电脑左侧的扬声器发送 18kHz 的正弦波，位于右侧的扬声器发送 19kHz 的正弦波，然后利用内置的麦克风以只的采样率对声音进行采样，由 Nyquist—Shannon 采样定理可知，采样率必须大于原始信号的 2 倍，因此在本实验中，我们让$f_{SB} = 48\text{kHz}$。大多数的笔记本电脑硬件都支持该采样率。在本实验的测试中，手势的最大频移几乎不会超过 200Hz。

在扬声器持续地发送高频声波信号的期间，任何在笔记本电脑附近的运动所产生的多普勒频移都会被麦克风记录下来，这在接收信号的频谱中表现为频率的增大或者减小。我们将麦克风收集到的声音信号通过抗混叠滤波器，保留$[16\text{kHz}, 20\text{kHz}]$的频率分量，并进行 6 倍欠采样，即$f_s = f_{SB}/6 = 8\text{kHz}$，把带通信号的频谱搬移到基带。然后以海明窗为窗函数进行$N_{FFT} = 4096$的短时傅里叶变换，每一帧选定的时长为$\delta_t = 85\text{ms}$，窗长$N_w = \delta_t f_s = 680$，窗之间的重叠长度为每个窗长的 75%. 由于离散傅里叶变换中频谱的对称性，每一帧都会得到等距离分布在 0Hz 到 4kHz 之间的 340 个频率点(frequency bin).

3.3 手势特征提取

在数据采集和预处理步骤完成后，我们得到了一系列的 FFT 频率矢量$F[k]$，我们需要通过手势特征提取模块从中提取出手势特征。

3.3.1 次峰搜索

当用户进行手势操作的时候，在此期间手势移动将会持续的产生多普勒频移，这些频移构成多普勒频移的时间序列，在本文定义的六个手势中，不同的手势将产生不同的频移时间序列，因此，通过匹配两个高频声音信号附近产生的多普勒频移时间序列，我们可以识别出用户做出了哪种手势，根据 2.1 节，手势移动相当于一个能够反射声波信号的移动平面。如图 1 所示，图中展示了频域中存在着幅值最大的两个峰，习惯上我们称之为主峰，这两个主峰是由扬声器和麦克风之间的视线距离(1ine of sight)传播路径和其他静止物体的反射叠加而成，因为在这些传播路径中波源和接收者的相对距离没有发生改变，所以这两个主峰的频率值与扬

声器发送的声波频率相同，即分别为 18kHz 和 19kHz，在图 1 中还可以看到依附于主峰左右两侧的次峰，其幅值均低于主峰，这些次峰的产生是因为附近存在着运动的物体，在本文中，就是由手势移动产生的频移，次峰可能位于主峰右侧，称作多普勒正移，也可能位于主峰左侧，称作多普勒负移，取决于手势移动的方向。但是，这些频谱图对于计算机而言是一串频率数据值，无法像人一样直观的看出次峰是在左侧还是右侧，为了让计算机找到次峰出现的位置，我们提出了一种次峰搜索方案，该方案可以比较准确的找出手势引起的多普勒频移的方向，具体过程如下：

Algorithm Secondary Peak Searching

Input: spectrum S of FFT length 4096
Output: Variable d indicating the direction with respect to the speaker. Value $+1$ for forward, -1 for backward and 0 for stationary.
$N \leftarrow$ the FFT point w.r.t. the frequency of the speaker.
$\Delta n \leftarrow$ spectrum resolution.
$L \leftarrow \{k|\, S[k] \geq \beta, N - \frac{100}{\Delta n} \leq k < N\}$
$l \leftarrow \max_{k \in L} S[k]$
$R \leftarrow \{k|\, S[k] \geq \beta, N < k \leq N + \frac{100}{\Delta n}\}$
$r \leftarrow \max_{k \in R} S[k]$
if $l > r$ **then**
    $d \leftarrow -1$
if $l < r$ **then**
    $d \leftarrow 1$
else
    $d \leftarrow 0$

其中$d = 1$表示次峰在主峰右侧，即多普勒正移，反之当$d = -1$表示次峰在主峰左侧，即多普勒负移。算法的主要思路为：以扬声器发送的声波频率值对应的 FFT 点N为中心，向左右两侧 100Hz 的频率范围扫描，寻找幅度超过阈值β的所有频点。将中心频率左右侧最大的幅度进行比较，从而确定次峰在主峰左侧还是右侧。阈值β与环境噪声和扬声器功率、扬声器与麦克风的远近有关。在室内条件下，本实验选取的$\beta = -20$dB. 如果两侧都没有搜索出幅度大于阈值的频率分量，则返回$d = 0$，代表没有次峰。

3.3.2 手势特征选取

在对每帧的频率矢量$F[k]$进行次峰搜索步骤之后，我们得到了手势频移的方向d，在前面中提到，我们通过发送两个高频声波信号，频率值分别为 18kHz 和 19kHz，因此经过次峰搜索步骤之后，我们应该在同一时刻得到两个频移值d，分别记为d_1和d_2. 我们将其放在一起作为一个二元组(d_1, d_2)，在用户进行手势操作期间，DoGest 将会持续的检测手势运动引起的多普勒频移，因此最终将会产生一个以二

元组为单位的时间序列，其形式类似于：

$$[(d_1(t_1), d_2(t_1)), (d_1(t_2), d_2(t_2)), \cdots]$$

经过综合考虑，我们将上述序列作为手势分类器的手势特征。此序列长度是不固定的。该序列长度取决于手势的移动速度，手势移动越快，手势持续时间就越短，该序列就越短，反之手势移动越慢，手势持续时间就越长，该序列越长，我们选取该序列作为手势特征的原因在于，运用此特征可以区分左右滑动手势，而这在单扬声器的情况下几乎是做不到的，在单扬声器单麦克风的情况下，无论是左滑手势还是右滑手势，都是手势先靠近麦克风，然后又远离麦克风，即先产生多普勒正移，然后又产生多普勒负移，对应的手势序列为：

$$[1,1,\cdots,-1,-1,\cdots]$$

由于这两个手势此时产生了相同的多普勒频移序列，我们显然无法从该序列中来区分用户是向左滑动还是向右滑动。而在我们的设计方案中，我们让两个扬声器分别发送不同频率的声波信号，可以获得更多关于手势移动的信息，从而可以区分出这两种手势。

为了便于解释，我们可以将笔记本电脑附近的区域划分为三个区域，第一个区域为以右侧扬声器为及其右边的区域，第二个区域为左右两个扬声器之间的区域，第三个区域为左侧扬声器及其左侧的区域，在用户向左滑动的过程中，我们把左滑动手势拆分为三个阶段，在第一个阶段中，手的运动过程为从右侧扬声器的右侧区域滑至两个扬声器之间的过程，该过程中，手势不断的靠近两个扬声器，然后远离右侧扬声器并继续靠近左侧扬声器，我们假设右侧扬声器的声波频率为19kHz，对应于二元组中的第二个多普勒频移方向，此时对应的时间序列为[(1,1)$^+$,(1,0)$^+$,(1,-1)$^+$]. 第二个阶段中，手的运动过程为在两个扬声器之间滑动的过程，该过程中，手势远离右侧扬声器，靠近左侧扬声器。此时对应的时间序列为[(1,-1)$^+$]，第三个阶段中，手的运动过程为从两个扬声器之间的区域滑动至左侧扬声器的左侧区域的过程，该过程中，手势远离右侧扬声器，靠近左侧扬声器，然后远离两个扬声器。此时对应的时间序列为[(1,-1)$^+$,(0,-1)$^+$,(-1,-1)$^+$]. 而在向右滑动的手势中，手势滑动过的每个阶段对应的时间序列都与左滑手势不一样，由此我们可以区分出这两个手势。需要说明的是，对于转圈手势，由于每个人的起始位置不同，旋转方向不同，不能用一个表达式表达，因此在表格 2 中我们没有显示写出转一圈和转三圈的时间序列。但是利用录制训练集和 kNN 分类，可以区分出转圈手势和其他四个手势。

表格 2 手势集

Gesture	Feature	Label
push	$[(1,1)^{+}]$	1
pull	$[(-1,-1)^{+}]$	2
swipe left	$[(1,1)^{+},(1,0)^{+},(1,-1)^{+},(0,-1)^{+},(-1,-1)^{+}]$	3
swipe right	$[(1,1)^{+},(0,1)^{+},(-1,1)^{+},(-1,0)^{+},(-1,-1)^{+}]$	4
…	…	…

3.4 手势分类

我们一共定义了六种手势，每种手势对应一个特定的时间序列正则表达式。手势分类器将手势序列识别为对应手势的标签值，在本文中，DoGest 系统利用 kNN 分类器来识别这六种手势，传统的 kNN 分类器利用欧几里得距离作为样本点之间距离衡量的尺度。然而，在我们的情景中，手势特征是时序的，意味着其序列长度是不定长的，其长度取决于手势移动的速度和手势移动的范围，考虑到以上情况，我们采用 2.5 节中提到的动态时间规整(DTW)方法计算样本之间的相似度并将其作为 kNN 分类器中样本点之间的距离度量。

因此，手势分类的大致流程为：

1）收集手势样本集，然后手工的给每个样本值添加相应的标签值

2）对手势进行识别，此时，用户进行手势操作，DoGest 利用次峰搜索算法找出手势引起的频移方向d，提取出时间序列作为待识别的手势特征，将此时间序列记为x，识别的过程为对带标签的所有样本值x_i，利用动态时间规整来计算$DTW(x, x_i)$作为待识别序列x与样本x_i的距离。找出距离最小的前k个样本记为$\{x_{i_1}, x_{i_2}, \cdots, x_{i_k}\}$，并通过“投票法”来作为待识别序列$x$的最终识别结果。

3.4.2 FastDTW 算法

FastDTW 算法使用减少搜索空间的策略来降低搜索复杂度[4]。时间复杂度 $0(m+n)$

(1) 粗粒度化亦即首先对原始的时间序列进行数据抽象，数据抽象可以迭代执行多次 1/1->1/2->1/4->1/16，粗粒度数据点是其对应的多个细粒度数据点的平均值。

(2) 投影。在较粗粒度上对时间序列运行 DTW 算法。

(3) 粒度细化。将在较粗粒度上得到的归整路径经过的方格进一步细粒度化到较细粒度的时间序列上。除了进行细粒度化之外，我们还额外的在较细粒度的空间内额外向外(横向，竖向，斜向)扩展 K 个粒度，K 为半径参数，一般取为 1

或者 2.

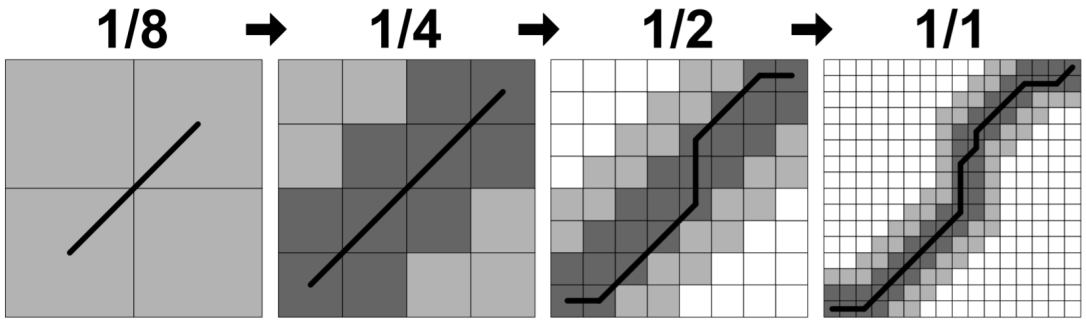


图 4 FastDTW 算法示意图

图 4 FastDTW 算法示意图中第一幅图表示在较粗粒度空间(1/8)内执行 DTW 算法。第二个图表示将较粗粒度空间(1/8)内求得的归整路径经过的方格细粒度化，并且向外(横向，竖向，斜向)扩展一个(由半径参数确定)细粒度单位后，再执行 DTW 得到的归整路径。第三个图和第四个图分别是在(1/2)和(1/1)粒度空间搜索的结果。

FastDTW 的伪代码如下

Algorithm FastDTW

1: **Input:** X – a Time Series of length $|X|$,
Y – a Time Series of length $|Y|$,
K – distance to search outside of the projected warp path from the previous resolution when refining the warp path.
2: **Output:** A min. distance warp path between X and Y,
The warped path distance between X and Y.
3: $MIN\_TS\_SIZE \leftarrow K + 2$
4: **if** $|X| \leq MIN\_TS\_SIZE$ *or* $|Y| \leq MIN\_TS\_SIZE$ **then**
return DTW(X, Y, $FullWindow$)
5: **else**
6: $shrunkX \leftarrow$ reduce by half of X
$shrunkY \leftarrow$ reduce by half of Y
7: [~, $lowResPath$] ← FastDTW($shrunkX, shrunkY, K$)
8: $window \leftarrow$ expanding the window using $lowResPath$ with X, Y, K
9: **return** DTW($X, Y, window$)

3.5 游戏接口与界面展示

为提高课程设计的趣味性与可展示性，我们设计了一款基于手势识别的推箱子游戏。该游戏基于 python 的 pygame 库编写。为调用基于 matlab 编写的多普勒效应识别模块，在基于 python 编写的游戏部分需使用 eng = matlab.engine.start_matlab()语句，在游戏主程序中即可调用 matlab 多普勒效应识别的结果。

目前为止，本游戏共有三个关卡，分别对应简单、中等、困难模式，图 5 展示了简单模式的界面。为充分运用多普勒效应可识别的六种手势，我们将 push, pull,

swipe left, swipe right 手势分别与向上、向下、向左、向右对应，single loop, triple loop 分别与撤回本步操作、从头开始游戏对应。该种对应方式获得了最好的游戏体验。值得说明的是，之所以用 triple loop 而不是 double loop 代表从头开始游戏，是为了将其与 single loop 更好区分开，以减少识别

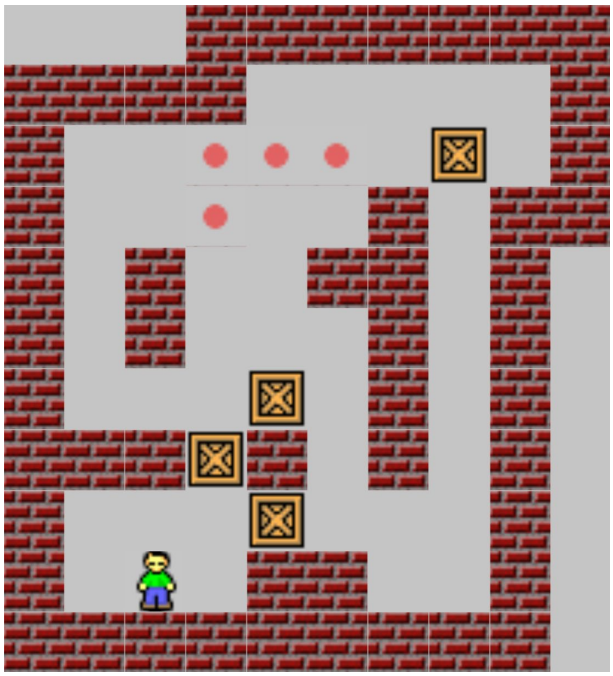


图 5 推箱子游戏简单模式界面

3.6 实验结果

3.6.1 手势识别准确度

为了测试 DoGest 的手势识别的准确率，我们一共邀请了五位参与者进行实验测试，具体实验过程为：首先，每个参与者每个手势执行 50 次，然后我们手动给每个取得的样本加上标签值，并将这些样本值作为我们的训练集，然后每个参与者再次执行每个手势 50 次，作为我们的测试集，因此，我们从每个参与者中收集到 50*2*6=600 个带标签值的样本，所有参与者总的样本数量为 600*5=3000 个，每个手势有对应的 500 个样本。我们将第一次执行的带标签值的手势样本集作为训练集，第二次收集的手势样本集作为测试集。表格 3 显示了所有测试者的平均准确率。

表格 3 手势分类混淆矩阵

push	0.99	0.00	0.00	0.01	0.01	0.01
pull	0.00	0.98	0.00	0.00	0.01	0.02
swipe left	0.00	0.00	0.94	0.03	0.04	0.02
swipe right	0.00	0.00	0.05	0.95	0.03	0.01
single loop	0.01	0.01	0.01	0.01	0.89	0.02
triple loop	0.00	0.01	0.00	0.00	0.02	0.92
	push	pull	swipe left	swipe right	single loop	triple loop

第四章 总结与展望

4.1 全文总结

对于人类来说，通过手势表达自身的想法是如此的自然，以至于我们在说话的时候经常会在不经意间做出手势比划，人类最早的交流方式是通过手势而非语言。凌空手势识别技术使得计算机理解人类的手势成为现实。

在本课程设计中，主要实现了基于双频的手势识别系统 DoGest，DoGest 通过利用两个不一样的声音频率识别手势，实现平台为笔记本电脑，且不需要任何其他额外的硬件设备。我们通过生成多普勒频移二元组的时间序列，并将其作为手势特征，然后使用动态时间规整(DTW)方法衡量手势特征之间的相似度，并使用 kNN 作为手势的分类器进行手势分类。DoGest 可以识别六种基本手势，包括上下滑动，左右滑动，转一圈，转三圈。经过多次测试，实验表明 DoGest 的平均手势识别准确率均在 94%以上。将手势识别结果作为推箱子游戏的控制方式，增添了数字信号处理课程设计的趣味性与可展示性。

DoGest 依然有一些不足的地方：手势识别的准确度极度的依赖于声音的强度，持续高强度的声音可能会对人的听力产生影响，同时也会增加设备的能量损耗；转圈手势的手势识别准确率不是很理想，有待提高；识别的手势集偏少，仅仅有六种，手势的数量有待提高。这些问题有待于通过提出新的手势特征方法解决。

4.2 技术展望

实际上，利用超声波进行位置定位已经有了比较成熟的解决方案，特别是在水下探测的场景中，因为电磁波信号在水中传输时能量衰减快而不适合作为定位的载波信号，而由于声波在水下的传输距离长等特性，人们已经开发出了非常成熟的声呐系统作为水下定位和目标检测的方案。而在陆地上，电磁波和声波的传输能力正好相反，GPS 球定位系统利用电磁波作为载波，通过确定信号的往返时间来测量物体的距离，一般应用于户外场景定位，比如汽车导航，智能手机定位，船只导航等。在室内场景中，由于房屋建筑对 GPS 信号的干扰使得 GPS 并不适合室内定位，且由于电磁波的速度非常大，这导致测量上极小的时间误差将会导致比较大的定位误差，因为 GPS 只有大约 10 米的精确度，所以在进行细粒度的定位时，一般不会对电磁波计算往返时间来测量距离，而比较多的方法是利用相位的方法进行距离估计，只需要计算出接收的反射电磁波和发送电磁波的相位差。

一般市面上的激光雷达的距离测量方法有很多就是基于相位法。激光雷达一般应用于自动驾驶和机器人定位与建图(Simultaneous Localization and Mapping)，其

成本一般来说都很高，不适合作为大众普及的个人使用的室内定位设备。声波相比电磁波拥有相对慢得多的传播速度，这意味我们可以利用计算传输的声波信号的飞行时间的方法来计算机物体的距离，而且大多终端设备都会配备扬声器和麦克风硬件，因此基于超声波的室内定位作为辅助定位具有一定的研究前景。

致　谢

在本学期数字信号处理课程的学习过程中，我们很荣幸能够得到武畅老师的教导。每次上数字信号处理课程都像是打开了新世界的大门，在本门课程的学习中，我们不仅收获了扎实的专业知识，还在课程学习过程中增强了人文素养与工程素养，使人终身受益。

本课程设计的工作是在武畅老师悉心指导下完成的，感谢他在课程设计的选题、实现、答辩过程中全方位的指导和宝贵意见。

参考文献

[1] FU B，KAROLUS J,GROSSE-PUPPENDAHL T．et a1．Opportunities for activity recognition using ultrasound doppler sensing on unmodified mobile phones[C]//Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction．ACM，20 1 5：8．

[2] QIFAN Y HAO T，XUEBING Z，et a1．Dolphin：Ultrasonic-based gesture recognition on smartphone platform[C]//2014 IEEE 17th International Conference on Computational Science and Engineering．2014：1461-1468．

[3] AI H，MEN YHAN L，et a1．High precision gesture sensing via quantitative characterization of the doppler effect[C]//2016 23rd International Conference on Pattern Recognition (ICPR)．IEEE，2016：973-978．

[4] Stan S，Philip C，FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space[C]// KDD Workshop on Mining Temporal and Sequential Data，2004：70-80．

[5] Vintsyuk，Taras K，Speech discrimination by dynamic programming[J]//Cybernetics and Systems Analysis 4.1，1968：52-57．