实验一 汉字操作

1.1 实验目的

本实验要求学生了解汉字字符编码的知识，包括 Unicode 符号集与典型实现 UTF8 编码，掌握 c# 编程语言条件下对汉字操作方法。通过项目学生应能向项目添加合适的引用文件，掌握编写简单程序的能力。

1.2 字符编码与实现

在计算机中，字符是以字节表示，美国对英文字符进行编码称为 ASCII 码，ASCII 码规定了 128 个字符编码；在处理非英语字符时，就需要使用其它的字符编码，汉字在计算机中的就有多种不同的编码方式，简体中文常见的编码方式是 GB2312，使用两个字节表示一个汉字，所以理论上最多可以表示 256x256=65536 个符号，中文编码还有 GB2312，GBK，BIG5 等。世界上存在着多种编码方式，同一个二进制数字可以被解释成不同的符号。打开文本文件时，就必须确定它的编码方式，否则用不正确的编码方式解读，就会出现乱码。在计算机应用程序中经常出现文本乱码现象，比如网页应用程序中，因为文本的制作和显示使用的编码方式不一样，这在一些应用程序中会显示一个小问号“？”，或者显示一个方框。Unicode 是一个符号集，它希望将世界上所有的符号都纳入其中，但是 Unicode 符号集在使用中会出现下面的几个情况：

1. 如果统一用等长的字节数来表示一个字节，则会产生浪费，例如使用四个字节来表示一个英文字符，则浪费了三个字节；2. 如何将 Unicode 与普通的 ASCII 码区别。Unicode 没有规定二进制存储方式。随着互联网发展，UTF-8 成为最流行的一种 Unicode 实现方式，另外还有

UTF-16 和 UTF-32 编码。而在.NET Framework 平台中，内存中的字符以 Unicode 进行编码。.NET Framework 提供 Encoding 类表示字符编码，它还可以方便将字符的编码互相转换。

1.3 计算机中汉字处理介绍

西方文字属拼读文字字符总数较少，中文是字形文字字符个数超过十万，虽然计算机处理西文非常容易但处理中文的过程却很复杂，涉及很多拼读文字没有的技术。例如多数人看到汉字不知道其读音，同个字符有几种注音。计算机对中文字符进行编码表示，由于复杂历史原因，汉字有中国大陆简化汉字的 GB2312 为编码方案，台湾及香港广泛使用的繁体字 BIG5 编码方案，有些程序要在繁体字与简体字之间进行转换。

为方便汉字字符处理，MS 提供了一个 Microsoft Visual Studio International Pack 类库软件包，目前是 1.0 SR1，从 MS 网站下载支这个类库软件包，下载完毕后，进行解压，里面有 7 个 msi 文件，这个软件包包含几个安装文件，其中 CHSPinYinConv.msi 是用来获取简体字读音，CHTCHSConv.msi 包用来执行繁体字转为简体字，安装这些软件包就可以使用其提供的类对汉字操作。

1.4 获取汉字的拼音

CHSPinYinConv.msi 软件包里面包含 Simplified Chinese Pin-Yin Conversion Library，它支持获取简体中文字符的常用属性比如拼音，多音字，同音字，笔画数。下面是获取汉字的拼音程序开发过程。

首先项目要添加引用库文件，应用 VS9 工具的菜单项目 –>添加引用，弹出添加引用对话框，参考图 1-1 切换标签页为浏览页，选择文件目标 ChnCharInfo.dll，例如 C:\Program Files\Microsoft Visual Studio International Pack\Simplified Chinese Pin-Yin Conversion Library\ChnCharInfo.dll。这个文件可拷贝到用户方便的目录中。

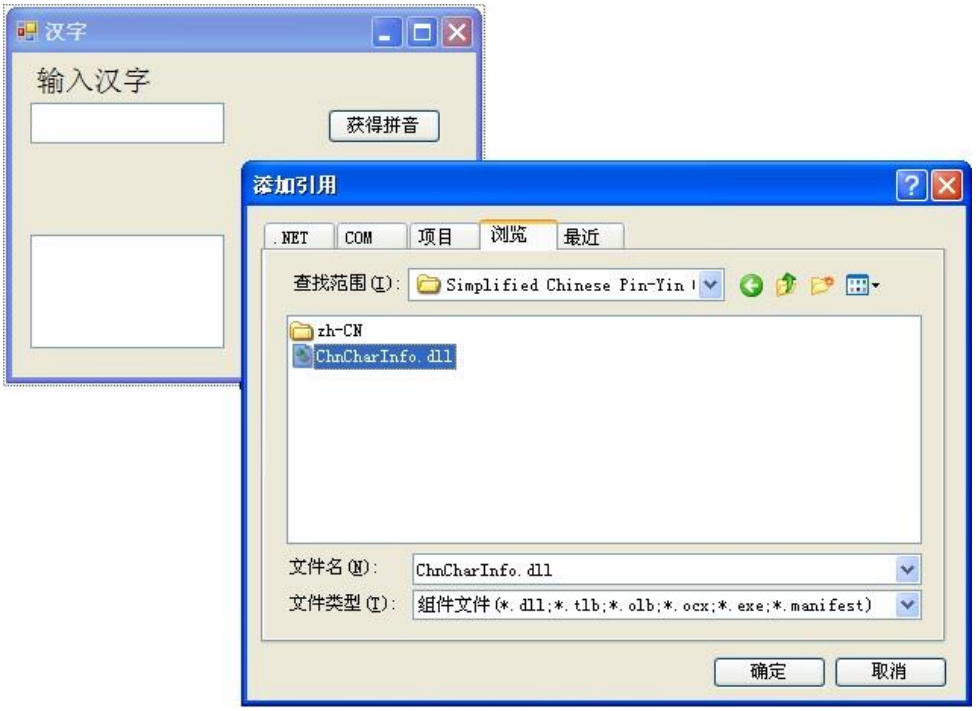

图 1-1 添加 ChnCharInfo.dll 引用

项目添加了引用文件后还要在程序上添加对应的命名空间。应用 VS9 工具菜单视图 –>代码进入代码编辑器，在源文件头中中添加 ChineseChar 类的命名空间：

```
using Microsoft.International.Converters.PinYinConverter;
using System.Collections.ObjectModel;
```

按钮 button1 的单击事件完成中汉字信息获取，添加下面的代码：

```
private void button1_Click(object sender, EventArgs e)
{ if (textBox2.Text.Trim().Length == 0)
  { return;
  }
  char one_char = textBox2.Text.Trim().ToCharArray()[0];
  int ch_int = (int)one_char;
  string str_char_int = string.Format("{0}", ch_int); if
  (ch_int > 127)
  {
    ChineseChar chineseChar = new
    ChineseChar(one_char); ReadOnlyCollection<string>
    pinyin = chineseChar.Pinyins; string pin_str = "";
    foreach (string pin in pinyin)
    { pin_str += pin + "\r\n";
    }
    textBox1.Text = "";
    textBox1.Text = pin_str;
  }
  label1.Text = str_char_int;
}
```

用户输入的字符作为字符串变量，它在.NET 平台中对应的是 string 类型，string 类型内置了丰富的函数方法，trim 方法用来将字符串首尾的空格符去掉，也可以通过参数指定要滤除的字符，它的返因结果仍是字符串变量。string 的 ToCharArray() 方法把当前字符串变量转化为 char 数组，以数组方式访问单个字符。此方法示例代码如下，它取得用户输入文本的首个字符。

char one_char = textBox2.Text.Trim().ToCharArray()[0];

中文存在一字多音现象，单个中文字符通过函数获取的拼音返回结果是 ReadOnlyCollection<string> 的数组形式。程序通过循环语句 foreach 将数组内的注音结果输出到多行文本框中，实现注音显示。注音结果的数字 1 至 4 表示汉字的四声，数字 5 表示轻声。

1.5 繁体字转换为简体字

安装软件包 CHTCHSConv.msi 后，程序库被安装到目录 C:\Program Files\Microsoft Visual Studio International Pack\Traditional Chinese to Simplified Chinese Conversion Library and Add-In Tool。下面是繁体字转换为简体字的程序实现，应用 VS9 工具向项目添加库文件的引用，方法同上小节相同，选择需要的库文件 ChineseConverter.dll 即可在程序中使用 ChineseConverter 类。应用菜单视图 –>代码进入代码编辑器，在源文件中添加命名空间：

using Microsoft.International.Converters.TraditionalChineseToSimplifiedConverter;

在按钮 button1 的单击事件将输入的繁体字转换为简体字，通过 ChineseConversionDirection 参数指定繁简转换的方向，例如输入" 北京時間"，则输出" 北京时间"，将代码中的 TraditionalToSimplified 改为 SimplifiedToTraditional 即可实现简体文本转换为繁体字的功能。示例代码如下：

```
private void button2_Click(object sender, EventArgs e)
{
    textBox2.Text=
    ChineseConverter.Convert(textBox1.Text,
    ChineseConversionDirection.SimplifiedToTraditional);
}
```

图 1-2 WindowsXP 语音选项

1.6 简体汉字 TTS--文本到语音

Text To Spee(TTS) 即"从文本到语音"让机器说话，实现文本发声阅读和将音频输出为 WAV 文件功能，在 windows xp 及后续版本中将此功能的编程接口实现称为 SAPI，它能直接将文本内容转换为语音。在 windows xp 的控制面版中找到语音选项如图 1-2，默认的语音选项是 Microsoft sam，即英文男声发音。要让机器实现中文发音需要下载安装微软 TTS5.1 语音引擎 (中文).msi 文件，成功安装以后语音选项会添加 Microsoft simplified Chinese 选项，本小节代码实现 TTS 功能就是调用 sapi.dll 库文件中相应函数。

使用 Windows 的查找命令找到文件 Tlbimp.exe 与 sapi.dll 文件，应用命令：

Tlbimp sapi.dll /out: DotNetSpeech.dll

它将原版的 sapi.dll 库文件转换为.NET 平台可用类，转换过程会输出一些信息用户可将其忽略。在程序窗体添加" 文本 ->发音" 按钮，将生成的 DotNetSpeech.dll 文件使用添加引用的方法加入到项目中，在程序代码最前面加上命名空间：

```
using DotNetSpeech；
```

添加按钮的事件代码：

```
private void button3_Click(object sender, EventArgs e)
{
  SpeechVoiceSpeakFlags spFlags = SpeechVoiceSpeakFlags.SVSFlagsAsync;
  SpVoice voice = new SpVoice();
  voice.Speak(textBox2.Text.Trim(), spFlags);
}
```

编译后运行程序，在文本框中输入一段话点击" 文本 ->发音" 按钮，会听到由计算读出的文本声音。TTS 中还提供了 SpeechStreamFileMode 类，它可将语音保存为 wav 文件，有兴趣的读者可尝试进一步应用

1.7 思考与练习

1. 在获取汉字拼音的代码中，将字符转化为数字的目的是什么？。
2. 以 UTF-8 编码保存汉字，所占的字节数目有什么特点，一个中文字符将由几个字节表示？
3. 实现简体文本转换为繁体字功能代码。