第 3 章　数字音频技术

学习目标

本章将重点介绍数字音频的基本概念和处理技术。

- 采样、量化、编码的概念。
- 数字音频压缩概念及常见文件格式。
- 常用音频编辑处理软件。
- 语音合成与语音识别的目的和作用。
- 使用 Soundbooth CS 实现音频文件的淡入淡出、去除噪音、调整语速、音量匹配、特效处理和多路音频合成等功能。

自然界中的声音是由于物体的振动产生的，是通过空气传播的一种连续的波，是模拟信号；数字媒体技术中的声音是经过计算机处理的，是数字音频。

3.1　数字音频基本概念

自然界的声音经过麦克风后，机械运动被转化为电信号。在计算机处理和存储声音之前，通过声卡把这些电信号转换为二进制数，这个转换过程称为模数转换。模数转换的过程分为采样和量化两个部分，经过处理后的音频信号就变成了数字音频。

对数字音频采用一定的算法进行压缩，然后采用一定的格式进行记录，这个过程称为编码。将编码后的数据存储在磁盘上，就形成不同格式的音频文件。

当用 Windows 的“附件”中的“录音机”打开某个音频文件，执行“文件”|“另存为”命令，在图 3-1 所示的“另存为”对话框中单击“更改”按钮，可以在打开的“声音选定”对话框中看到采样频率、量化位数和编码等信息，如图 3-2 所示。

注意

- 采样频率越高，量化位数越多，记录的波形越接近原始信号，声音的质量就越高，当然，所占用的存储空间也越大。
- 目前常用的标准采样频率是 8kHz、11.025kHz、22.05kHz、44.1kHz 和 48kHz，常用的量化位数为 8 位和 16 位。
- CD 音质的采样频率为 44.1kHz，量化位数是 16 位。

声道数是指所使用的声音通道的个数，它表明声音记录是产生一个波形（即单音或单声道）还是两个波形（即立体声或双声道）。虽然立体声听起来要比单音丰满优美，但需要两倍于单音的存储空间。

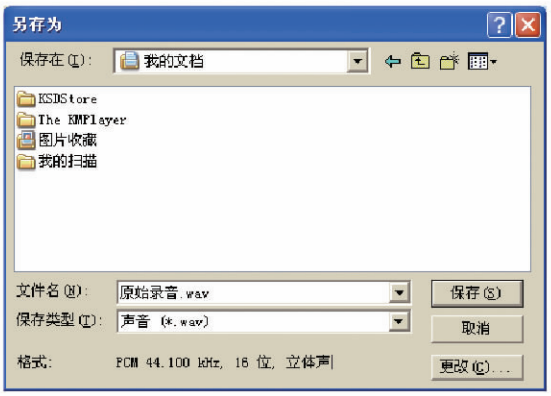

图 3-1 “另存为”对话框

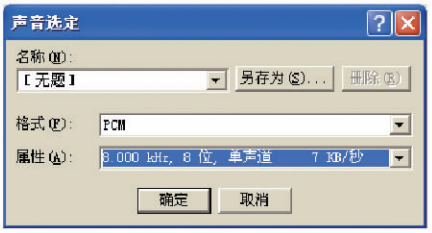

图 3-2 采样频率、量化位数和编码

采样频率、量化位数和声道数对声音的音质和占用的存储空间起着决定性的作用。存储容量与上述三要素之间的关系可用下列公式表示：

$$存储容量(byte)=\frac{采样频率(Hz)\times量化位数(bit)\times声道数}{8}$$

一般来说，希望音质越高越好，磁盘存储空间越少越好，这本身就是一个矛盾，必须在音质和磁盘存储空间之间取得平衡。在数字媒体项目开发与制作中，声音文件一般推荐22.050kHz、16 位。它的数据量是 44.1kHz 声音的一半，但音质却很相似。

将经过采样和量化的数字音频不进行压缩，直接记录下来而形成的文件格式是 PCM。PCM 的最大优点是音质好，最大缺点是文件大。

注意

- 常见的 WAV 文件就是微软公司开发的一种基于 PCM 编码的声音文件格式。
- Audio CD 也采用 PCM 编码，一张光盘只能容纳 72 分钟的音乐信息。

3.2 数字音频压缩技术

未经压缩的数字音频占据了大量的存储空间，以 CD 为例，其采样频率为 44.1kHz，量化位数为 16 位，1 分钟的立体声音频信号需占约 10M 字节的存储容量，也就是说，一张 CD 唱片的容量只有 1 小时左右。事实上，在无损条件下对声音至少可进行 4∶1 的压缩，即只用 25%的数字量保留所有的信息，而在视频领域压缩比甚至可以达到几百倍。因而，为了利用有限的资源，压缩技术从一出现便受到广泛的重视。

数字音频压缩是指应用适当的数字信号处理技术，在不损失有用信息量或所引入损失可忽略的前提下，对原始数字音频信号(PCM 编码)进行压缩。

注意

由于数字音频压缩技术具有广阔的应用范围和良好的市场前景，因此一些著名的研究机构和大公司都不遗余力地开发自己的专利技术和产品，形成了各种压缩文件格式，如 MP3、RealAudio、WMA、ASF 和 OGG 等。

1. MP3 文件格式

MP3 是 MPEG(Moving Picture Experts Group)Audio Layer-3 的简称，是 MPEG1 的衍生编码方案，可以做到 12∶1 的惊人压缩比并保持基本可听的音质。在 MP3 中使用了知觉音频编码和心理声学，也就是利用了人耳的特性，削减音乐中人耳听不到的成分，以确定音频的哪一部分可以丢弃，同时尽可能地维持原来的声音质量。

2. RealAudio 文件格式

RealAudio(RA)是 Real Networks 公司开发的网络流媒体音频格式，特点是可以在非常低的带宽下(低至 28.8kbps)提供足够好的音质，并可以随着网络带宽的不同而改变声音的质量，在保证大多数人听到流畅声音的前提下，令带宽较宽敞的听众获得较好的音质。

3. WMA 文件格式

WMA(Windows Media Audio)是微软公司开发的数字音频压缩格式，其压缩率一般可以达到 18∶1。WMA 比 MP3 的体积少 1/3 左右，并支持网络流式播放。此外，WMA 支持证书加密，通过 Windows Media Rights Manager 加入保护，可以限制播放时间、播放次数和播放机器。

4. ASF 文件格式

ASF(Audio Steaming Format)除支持音频外，还支持视频及其他多媒体类型。由于 Microsoft 公司的大力推广，这种格式在高音质领域直逼 MP3，并且压缩速度比 MP3 提高 1 倍，在网络广播方面可与 RealAudio 相竞争。

5. OGG 文件格式

OGG(Ogg Vorbis)是一个完全免费、开放和没有专利限制的自由编解码器。在压缩技术上，Ogg Vorbis 最主要的特点是使用了 VBR(可变比特率)和 ABR(平均比特率)方式进行编码，与 MP3 的 CBR(固定比特率)相比可以达到更好的音质。

3.3 数字音频处理技术

数字音频处理包括音频采集、音频编码/解码、音频提取、格式转换、语音合成、语音识别、音频数据传输、音频视频同步、音频效果与编辑等。

音频编辑处理软件可以实现数字音频录制、混合、编辑、转换和制作等功能，如 Adobe 的 Soundbooth 和 Audition(Cool Edit)、GoldWave、WaveCN 等。

Audition 偏重纯音频的专业化处理，对音频进行复杂的处理；Soundbooth 将重点放在与视频的结合上，更像视频后期的配音平台，适合于处理视频、动画、网页中的声音。

GoldWave 和 WaveCN 都是集声音编辑、播放、录制、转换的音频工具，体积小巧，内含丰富的音频处理特效。WaveCN 还是一款免费的中文录音编辑处理软件。

语音识别和语音合成技术是实现人机语音通信的两项关键技术，语音合成技术的目的是使计算机具有说话能力，语音识别技术的目的是使计算机听懂人说的话。

3.4 语音合成技术

语音合成技术(Text To Speech，TTS)是利用计算机按人们预定的程序和指令，人为地产生出音素、音节、词和句子的技术。主要处理如何将文字信息转化为语音信息，以实现动

态的、及时的语音朗读等功能，涉及声学、语言学、数字信号处理技术、数字媒体技术等多个学科。

语音合成技术解决的主要问题就是如何将文本状态的文字信息转化为可听的声音信息，使以往只能用眼睛看的文字信息也可以用耳朵来听。

科大讯飞、NeoSpeech 和 ScanSoft 都处于业界领先的地位，其语音合成的音质和音感都非常出色，合成的语音自然度超过普通人说话的水平。其中，科大讯飞被国内外专家公认为具有世界最高水平的汉语语音合成技术。

【例 3-1】 应用科大讯飞“文语通”实现语音合成。

(1) 运行科大讯飞“文语通”，程序界面如图 3-3 所示。

(2) 单击“打开”按钮，在“打开”对话框中选择要播放的 htm、html、txt 或 rtf 格式的文件，即可实现朗读功能。

(3) “文语通”安装后，已经作为插件嵌入 Word 和 IE 软件中，实现在线播放的功能。

(4) 在 Word 中，单击图 3-4 所示的“朗读全文”或“朗读选定”按钮，可以对全文或选定部分朗读。

图 3-3 “文语通”程序界面

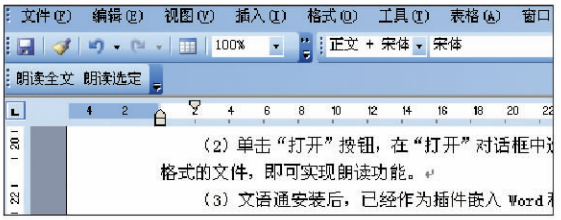

图 3-4 嵌入 Word 中的插件

(5) 浏览网页时，右击鼠标，从弹出的快捷菜单中选择“使用‘文语通’朗读选定内容”或“使用‘文语通’朗读链接”命令，如图 3-5 所示，即可实现朗读功能。

图 3-5 嵌入浏览器中的插件

【例 3-2】 应用微软 TTSAPP 将朗读文本输出为 WAV 文件。

(1) 运行 TTSAPP，程序界面如图 3-6 所示。

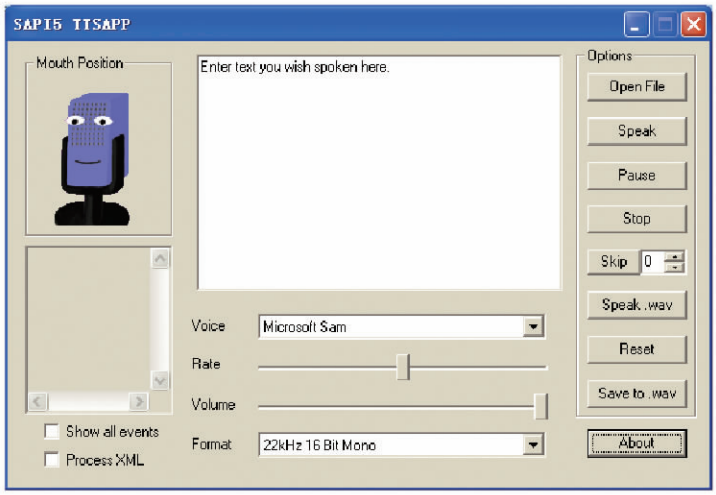

图 3-6 TTSAPP 的程序界面

(2) 单击 Open File 按钮选择文本文件，或直接在图中的文本框内输入文字。根据要输出的声音文件的质量选择 Format 下拉列表中的采样频率、量化位数和声道数；在 Voice 下拉列表中选择下载的高质量音库。单击 Saveto.wav 按钮，将文本合成为音频文件，如图 3-7 所示。

图 3-7 合成 WAV 音频文件

3.5 语音识别技术

语音识别技术是实现人机语音通信所必需的另一项关键技术，它的目的是使计算机具有听懂人说话的能力。

语音识别技术所涉及的领域包括信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等。

语音识别分为训练和识别两个阶段：训练阶段在机器中建立被识别语音的模型库；识别阶段将被识别的语音特征参量提取出来进行模式匹配，相似度最大者即为被识别语音，如图 3-8 所示。

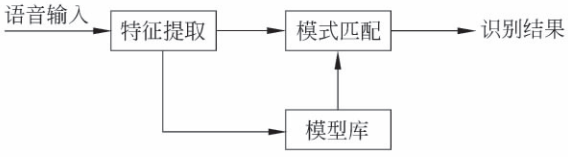

图 3-8 语音识别原理框图

语音识别技术的应用包括语音拨号、语音导航、室内设备控制、语音文档检索和简单的听写数据录入等。语音识别技术与其他自然语言处理技术如机器翻译及语音合成技术相结合，可以构建出更加复杂的应用，例如语音到语音的翻译。

较好的语音识别软件有 IBM 的 Via Voice、微软的语音识别系统 Speech SDK、Dragon

的 Naturally Speaking 等，Google 在 2010 年利用语音识别技术为 YouTube 视频自动生成字幕。

【例 3-3】 应用微软语音识别系统 Speech SDK 实现语音输入。

（1）运行控制面板中的“区域和语言选项”选项，在出现的图 3-9 所示的“区域和语言选项”对话框中单击“详细信息”按钮，出现“文字服务和输入语言”对话框，如图 3-10 所示。

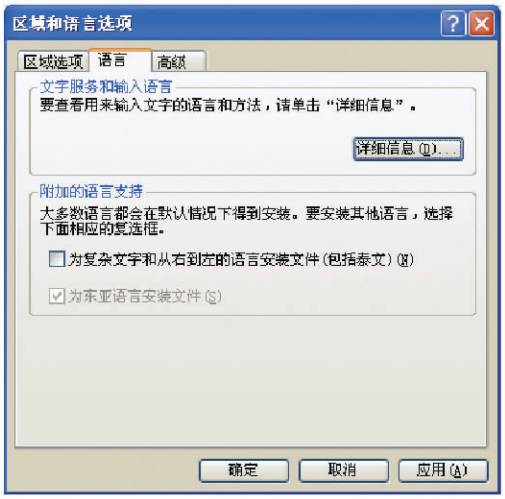

图 3-9 “区域和语言选项”对话框

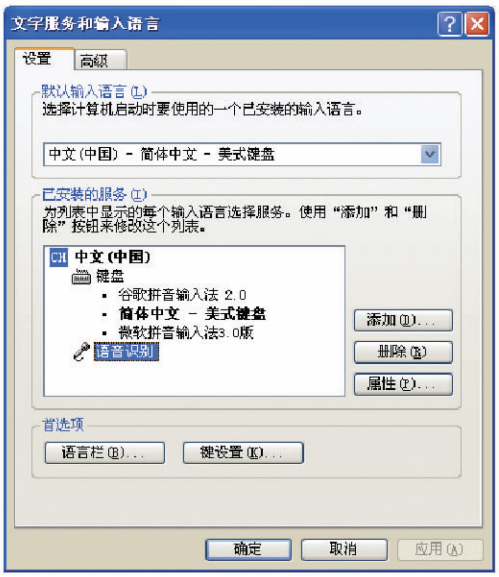

图 3-10 “文字服务和输入语言”对话框

（2）选择“语音识别”，单击“属性”按钮，出现“语音输入设置”对话框，如图 3-11 所示。

（3）单击“高级语音”按钮，出现“语音属性”对话框，如图 3-12 所示。

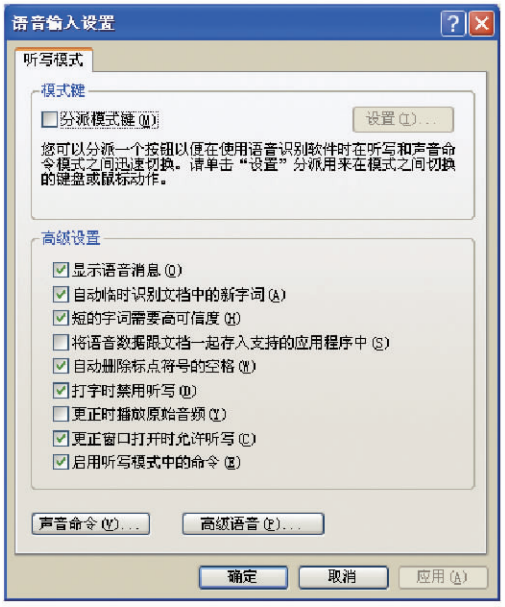

图 3-11 “语音输入设置”对话框

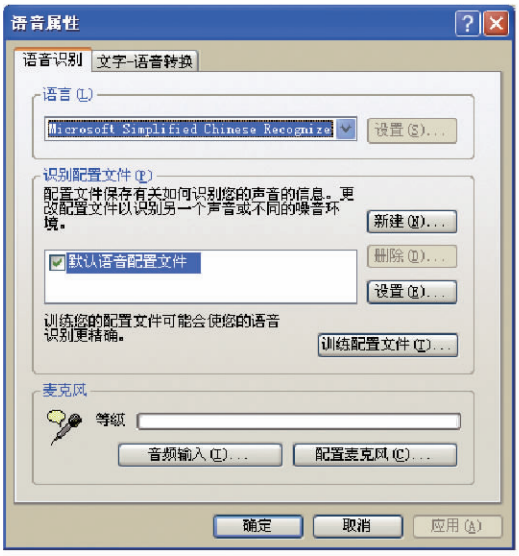

图 3-12 “语音属性”对话框

(4) 单击“配置麦克风”按钮，出现“麦克风向导”对话框，如图 3-13 所示。

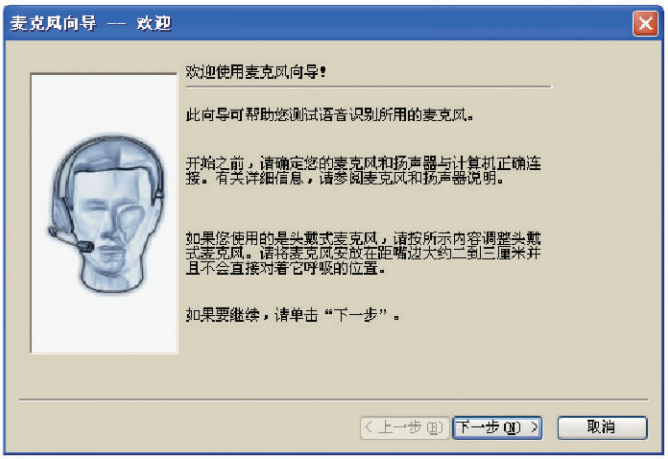

图 3-13 “麦克风向导”对话框

(5) 单击“下一步”按钮，出现“测试麦克风”对话框，测试麦克风，如图 3-14 所示。

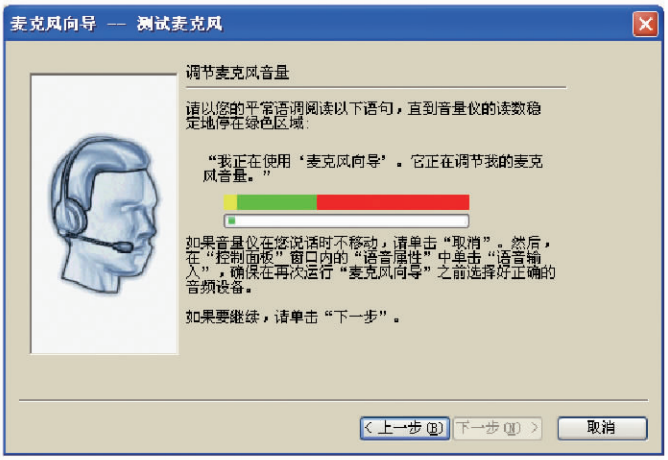

图 3-14 测试麦克风

(6) 按提示信息进行朗读，单击“下一步”按钮，出现“测试安放位置”对话框，如图 3-15 所示。

(7) 继续按提示信息进行朗读，听到正常的回放录音后，单击“完成”按钮，返回到图 3-12 所示的“语音属性”对话框。

(8) 单击“训练配置文件”按钮，出现“声音训练”对话框，如图 3-16 所示。

(9) 单击“下一步”按钮，进行声音训练，如图 3-17 所示。

(10) 朗读完成后，软件自动更新配置文件，如图 3-18 所示。

(11) 更新配置文件后，出现图 3-19 所示对话框，可以单击“更多训练”按钮，进行更多训练，或单击“完成”按钮，完成训练。

(12) 打开 Word，将输入法切换至微软拼音输入法，单击图 3-20 所示语言栏中的“麦克风”按钮，切换到听写模式，如图 3-21 所示。

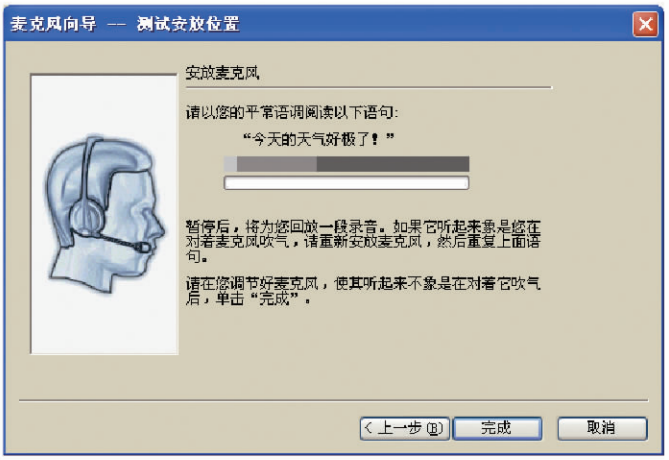

图 3-15　测试安放位置

图 3-16　“声音训练”对话框

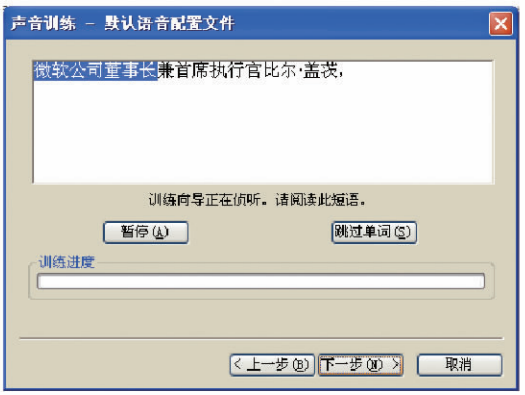

图 3-17　进行声音训练

图 3-18　更新配置文件

图 3-19　完成训练

图 3-20　语言栏

图 3-21　听写模式

注意

- 当进行朗读时，Word 就会将其自动转换为文字。
- 如果转换的文字错误较多，则需按上述方法进行多次声音训练。

3.6 音频处理软件 Soundbooth CS4

Adobe Soundbooth CS4 的设计目标是为网页设计和影像编辑人员提供高质量的声音信号，实现快速录音、编辑等功能。其主要优势是在和 Adobe 其他工具的配合上，可直接编辑 Adobe Premiere、Flash 和 After Effects 的音频格式。与以前版本相比，其最大的改变是开始支持多轨录音。

3.6.1 音频录制与格式转换

在工作区中，可以非常方便地完成音频录制工作。

【例 3-4】 音频录制。

(1) 运行 Soundbooth CS4，程序界面如图 3-22 所示。

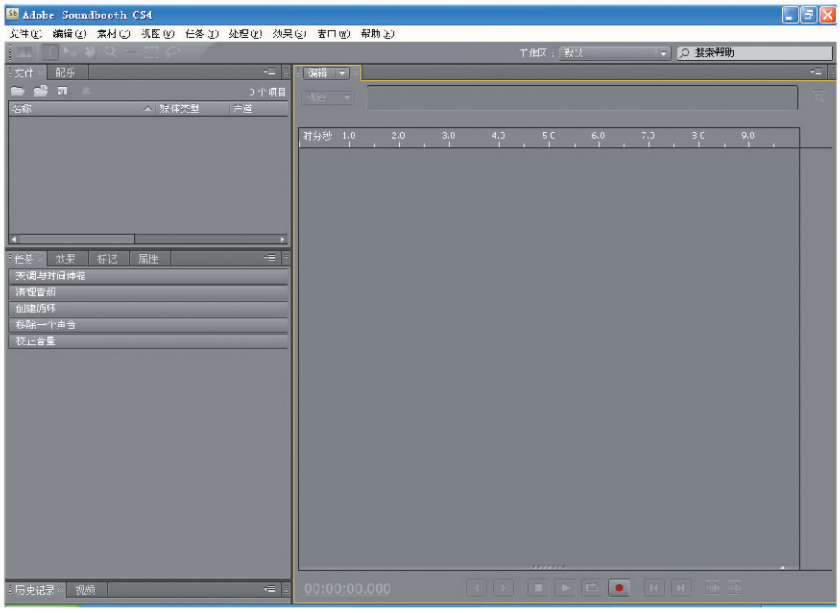

图 3-22 Soundbooth CS4 程序界面

(2) 单击右下角的红色录音按钮，出现“录音”对话框，如图 3-23 所示。

(3) 在对话框中选择相应参数，如“采样率”、“声道”等，单击左下角的红色“录音”按钮开始录音。录音完毕单击“停止录音”按钮停止录音，单击“关闭”按钮返回主界面。单击工作区中的播放按钮可以播放录制的声音文件，如图 3-24 所示，默认文件格式为 WAV 文件。

(4) 执行“文件”|“另存为”命令，出现“另存为”对话框，在“保存类型”下拉列表中可以选择音频文件类型，如图 3-25 所示。

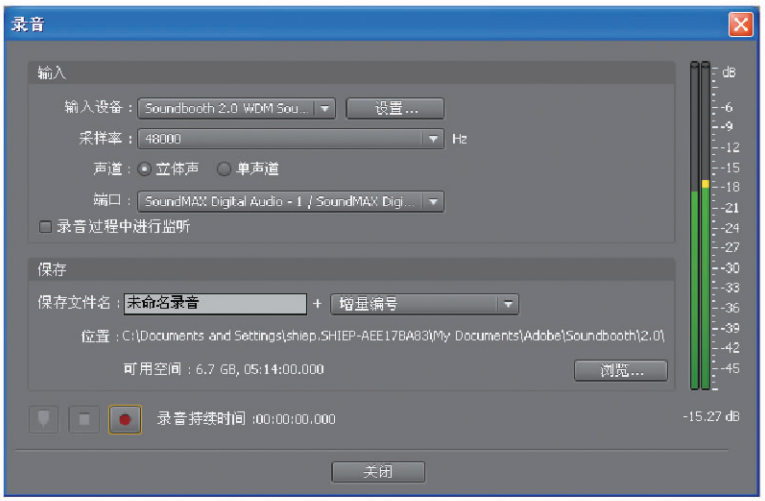

图 3-23 “录音”对话框

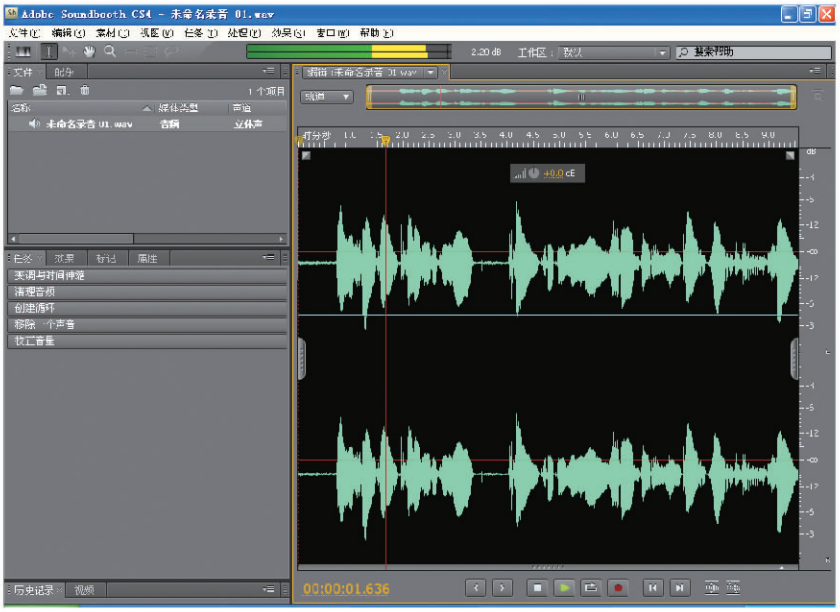

图 3-24 播放录音

图 3-25 保存文件