第 5 章

音 频 数 据


视频讲解

作为信息传递的重要载体，音频数据不仅承载着丰富的声音信息，还深刻影响着我们的感知与理解。从细微的耳语到宏大的交响乐，音频数据以其独特的方式记录并再现了世界的多样性和动态变化。本章旨在全面剖析音频数据的基本概念、特点、预处理方法及表示方式，为理解音频处理、分析及应用技术奠定坚实基础。

随着科技的飞速发展，音频数据的采集、处理与应用技术也在不断革新，为我们带来了更加丰富、细腻的声音体验。在本章的开始，我们将首先揭开音频数据的神秘面纱，探讨其基本概念与特性，为后续深入学习音频处理、分析及应用技术做好铺垫。

通过本章的学习，读者将对音频数据的基本概念、特点和处理方法有更深入的了解，并掌握处理和分析音频数据的关键技术。这对于在音频信号处理、音频信息检索、语音识别等领域的研究和应用具有重要的学术和实践意义。

学习目标

当完成本章学习后，要求：

- 了解音频数据的基本特点。
- 掌握音频数据清理方法。

学习难点

- 掌握音频数据分帧处理方法。
- 掌握音频数据离散表示方法。
- 掌握音频数据的分布表示方法。

5.1 音频数据基本特点

本节将介绍音频数据的基本特点，包括其连续性和时序性。连续性意味着音频信号在时间上的连贯展现，保证了声音的自然流动和完整性。时序性涉及音频数据的顺序性，对于正确解析音频信息至关重要。了解音频数据的连续性和时序性对于有效处理和分析音频数据至关重要。

5.1.1 连续性

音频信号作为一种随时间连续变化的实时一维信号，其本质特性赋予了音频数据在时域上的强烈连续性。在数字化转换过程中，尽管音频信号被离散化为采样点，但这些采样点在时间轴上依然维持着高度的连续性，这一点对于保留音频信号的完整性和真实性至关重要。在数字音频处理的范畴内，这种连续性体现在音频样本序列之中，成为影响音频质量与细节表现的关键因素之一。音频数据在物理样本

层、声学特征层以及语义层上，均展现出不同程度的连续性，这种连续性构成了音频信号处理与分析的基石。

如图 5-1 所示，在物理样本层面上，音频数据的连续性是通过采样过程来体现的。采样是对连续的模拟音频信号以固定的时间间隔进行测量，从而获取一系列离散的样本值。尽管这些样本值在时间上是离散的，但由于采样频率的选择（通常远高于人耳的分辨能力），它们能够极为逼真地模拟原始音频信号的连续性。采样频率决定了单位时间内的采样次数，采样频率越高，相邻样本之间的时间间隔就越小，所得到的数字化音频信号就越逼近原始的模拟音频信号，从而在物理样本层面上维持较高的连续性。以 CD 音质为例，其采样频率通常为 44.1kHz，即每秒对音频信号进行 44 100 次采样，这一采样频率足以确保音频信号在物理样本层面上的连续性。

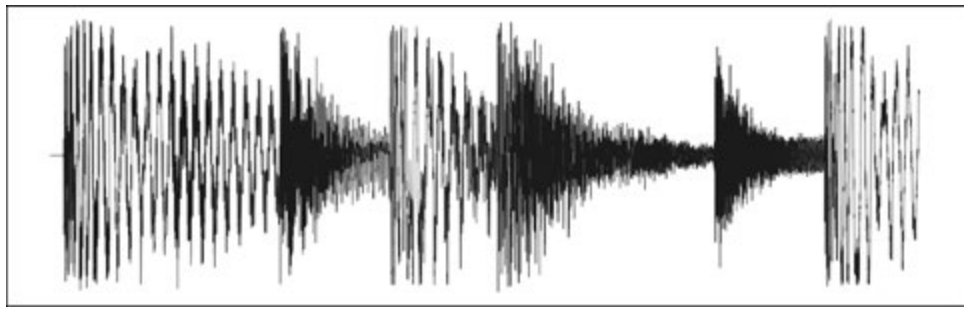

图 5-1　模拟的音频信号

在声学特征级上，音频数据的连续性主要体现在其感知特征上，包括音调、音高、旋律、节奏等。这些感知特征在音频信号中呈现出连续的变化，充分反映了声音的基本属性及其动态变化过程。通过数字信号处理技术，可以提取并分析这些感知特征，从而在声学特征级上实现对音频数据的连续性分析和处理。具体而言，通过分析样本数据中的短时能量和频率分布，能够深入理解音频的节奏和旋律变化，进而维持音频内容的连续性和表达的连贯性。

在语义级上，音频数据的连续性主要体现在其语义内容的连续性和完整性方面。音频数据可能涵盖语音、音乐或其他声音类型，这些声音类型都蕴含着特定的语义内容，如语音中的词汇、句子，以及音乐中的旋律、和弦等。这些语义内容在音频信号中是连续组织和表达的，共同构成了音频数据的整体意义。在语义级上，语音识别和音乐分析等技术均考虑了音频数据的连续性。语音识别技术能够准确识别并理解音频信号中的语音内容，并将其转换为文本形式。音乐分析技术能够提取和分析音频信号中的音乐特征，如旋律、节奏、和声等。这些技术不仅需要处理音频数据的物理和声学属性，还需要深入理解音频所传达的深层含义，以确保信息的连续传递和逻辑上的紧密衔接。

综上所述，音频数据的连续性确保了声音作为一种信息载体在时间和内容上的连贯性，这使得音频既能精确地表达原始的声音信号，也能在更高层次上传递复杂的信息和情感。

5.1.2　时序性

音频数据的时序性对音频信号的处理和分析至关重要，时序性贯穿音频信号的整个生命周期，从采集、传输到最终分析，每一步都无法脱离其时间属性的影响。音频处理常常依赖于前后时间点的数据，每个样本点的时间标记使得可以精确地定位到声音事件发生的具体时间，这对于同步视频与音频、音乐节拍检测以及动态处理音频效果等应用至关重要。

在时域分析中，如回声消除、噪声抑制和动态范围压缩等技术都直接处理时间序列数据，这些技术依赖于音频信号在时间上的特性。音频处理中常见的特征，如能量、节奏、音高和时长等，都是从时间序列

中提取的，这些特征的计算通常需要考虑连续时间窗口内的样本数据。

此外，音频数据的时序性在音频语义特征学习上有着重要的作用。如图 5-2 所示，时序关系依赖的音频语义特征学习过程主要由两个步骤构成：

（1）局部特征的提取，即基元表示的构建；

（2）局部特征之间的时序关系建模，即基元间时序信息的获取[1]。

图 5-2 时序关系依赖的音频语义特征学习过程

时序性不仅是音频理解的基础，还是多种音频处理技术的关键。在语音识别技术中，时序性是利用语音模型预测字词序列的基础。通过维持音频样本点的时间顺序，语音识别系统能够理解语境和语义的连续性，从而提高识别的准确性。同样，在音乐信息检索中，保持音频数据的时间顺序可以帮助学习算法更好地学习音乐的节奏和动态变化，从而实现更为精确的音乐分类和推荐。

在音频分析和音频处理领域，时序性确保了音频数据能够按照其原始的时间顺序进行分析和解释，这对于从语音识别到音乐信息检索的各种应用都是必不可少的。同时，保持音频的时序性不仅有助于提高识别和解析的准确性，还能增强对音频内容的整体理解。

5.2 音频数据预处理

在处理音频数据之前，预处理是一个关键的步骤。本节将详细讨论音频数据预处理的各个方面。其中，音频清洗是一项重要的预处理步骤，旨在去除噪声、干扰和其他不必要的信号，以提高音频数据的质量和可用性。此外，我们还将介绍采样频率和位宽转换的概念和方法，这对于音频数据的格式转换和适应不同应用场景非常重要。接下来，我们将介绍分帧处理和时频分析的技术，这些方法将音频数据分割成小的时间片段，并分析每个时间片段的频谱特征，以便更好地理解和处理音频数据。这些技术对于音频信号处理和语音识别等应用具有重要意义。图 5-3 是音频数据预处理流程。

图 5-3 音频数据预处理流程

5.2.1 音频清洗

音频清洗是音频处理中的一个重要环节，旨在提升音频数据的质量，使其更适合后续的分析、处理或应用。音频清洗是指对采集到的原始音频数据进行一系列的处理和修正过程，以消除或减少其中的噪声、干扰、失真等问题，同时保持或提升音频的清晰度、连贯性和准确性。音频清洗涉及多种技术和方法，包括但不限于如下方面。

1. 降噪和去除干扰

音频数据常常受到环境噪声、电磁干扰等影响，降低了数据的质量。通过使用自适应滤波、频谱减法和小波变换等数字信号处理技术，可以对音频数据进行降噪处理，去除环境噪声、电源噪声等干扰信号，以提高音频的清晰度和可听性。

2. 去混响

在一些场景中，音频数据可能受到混响的影响，使得声音产生回响，影响了数据的质量和可用性。在清洗音频数据的过程中，可通过线性预测编码、时频分析和波束形成等信号处理方法模拟和去除混响效应，以提高音频的清晰度和可理解性。

3. 端点检测

音频数据中可能包含了非语音部分，如静音或者其他噪声。在清洗音频数据的过程中，可以进行端点检测，确定语音的起始和结束位置，从而剔除非语音部分，提取出有效的语音段落，这有助于完成后续的语音识别和分析任务[1]。

4. 音量归一化

音频数据中的音量水平可能不一致，影响了数据的可用性和处理效果。通过峰值归一化、均方根归一化和伽马归一化等音量归一化处理方法，可以调整音频数据的音量水平，使其保持一致，以便更好地适应后续处理和分析任务。

5. 异常检测和修复

在音频数据中可能存在一些异常或损坏的部分，如断裂、重叠、失真等。在清洗音频数据的过程中，可以进行异常检测，并采取相应的修复措施，如插值、平滑等，以提高数据的完整性和可用性。

5.2.2 采样频率和位宽转换

音频数据的采样频率和位宽转换是预处理中的两个重要方面，它们直接影响音频信号的质量和特性。采样频率是指每秒钟采样数据的次数，用赫兹(Hz)表示。更高的采样频率意味着对声音信号进行了更频繁的采样，因此可以更准确地表示原始声音信号。根据采样定理[2]，采样频率大于信号两倍带宽时，采样过程不会丢失信息，并且根据采样信号可以精确地重构原信号。采样频率转换是指将音频信号的采样频率从一个值更改为另一个值的过程。这种转换通常用于不同采样频率要求的设备或系统之间的音频数据交换，或者为了优化音频信号的质量和处理效率。采样频率转换可以通过重采样技术实现，即根据原始采样频率重新计算音频信号的样本值，以匹配目标采样频率。重采样过程中需要采用适当的插值算法，以确保转换后的音频信号在听感上尽可能接近原始信号。

音频数据以数字形式存储时，每个采样点都被表示为一个数字值。位宽是指每个采样点表示音频信号幅度所用的比特数。常见的位宽有 8 位、16 位、24 位和 32 位等。例如，一个 16 位的音频数据表示每个采样点用 16 个二进制位来表示，因此可以有 2^{16} 种不同的取值，即 0～65 535。位宽决定了音频数据的动态范围和精度。较高的位宽可以提供更高的动态范围和更精确的表示，然而，较高的位宽也会导致音频数据文件更大，因为每个采样点需要更多的位数来表示。

位宽转换是指改变音频数据的位深度，即每个采样点所占的二进制位数。这种转换通常用于改善音频信号的动态范围、信噪比或与其他系统兼容。位宽转换可以通过位深度(位宽)缩放技术实现，即根据原始位宽重新量化音频信号的样本值，以匹配目标位宽。在转换过程中，需要保持音频信号的总体动态范围和重要细节，同时尽量减少量化噪声的引入。位宽转换可以分为两种情况。第一种是降低位深度，即将较高位深度的音频数据转换为较低位深度的数据。例如，从 24 位转换为 16 位或从 16 位转换为 8 位。这种转换会导致一定的信息丢失，因为较低的位深度无法精确地表示原始较高位深度的数据。第二种是提高位深度，即将较低位深度的音频数据转换为较高位深度的数据。例如，从 8 位转换为 16 位或从 16 位转换为 24 位。这种转换会增加数据的精度，但也可能会在某些情况下增加量化噪声。

在进行位宽转换时，需要考虑到以下几个方面的问题。位宽转换会引入量化误差，因为较低位深度无法完全表示原始较高位深度的数据。因此，在转换过程中需要采取一定的技术来减小量化误差的影响。此外，位宽转换可能会影响音频数据的动态范围，即音频信号的最大振幅和最小振幅之间的范围。

降低位深度会降低动态范围，而提高位深度会增加动态范围。

5.2.3 分帧处理

音频数据的分帧处理是指将连续的音频信号按照一定的时间长度分割成多个短时段的信号段，每个信号段称为一帧。这样做的目的是为了将非稳态的音频信号在短时段内视为稳态，从而便于后续的分析和处理。由于音频信号是长时间非稳态的序列，直接处理整个音频信号会非常复杂且难以提取有效特征，因此分帧处理成了一个必要的步骤[3]。原始样本数据往往包含着尖锐噪声，若不加以适当的滤波处理往往会影响分析效果。实践应用中一般采用一阶数字滤波来实现，即

$$H(Z)=1-\mu z^{-1} \tag{5-1}$$

其中，z 表示输入信号，μ 值接近 1。

进行过数字滤波处理后，可对数据进行加窗分帧处理，一般每秒为 33～100 帧，具体视实际情况而定。分帧虽然可以采用连续分段的方法，但一般采用交叠分段的方法，这是为了使帧与帧之间能够平滑过渡，保持其连续性。前一帧和后一帧的交叠部分称为帧移。帧移与帧长的比值一般取为(0,0.5)。分帧是用可移动的有限长度窗口进行加权的方法来实现的，就是用窗函数 $\omega(n)$ 来乘以原始语音信号 $s(n)$，从而形成加窗语音信号 $s_\omega(n)=s(n)*\omega(n)$，其中，$n$ 表示样本索引。常用的窗函数有矩形窗、海宁窗和汉明窗，它们的表达式如下(其中，N 为帧长)：

矩形窗

$$\omega(n)=\begin{cases}1, & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{5-2}$$

海宁窗

$$\omega(n)=\begin{cases}0.5(1-\cos(2\pi n/(N-1))), & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{5-3}$$

汉明窗

$$\omega(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)], & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{5-4}$$

不同窗函数的具体特性如表 5-1 所示。

表 5-1 不同窗函数的具体特性

特性	矩形窗	海宁窗	汉明窗
A	13dB	32dB	43dB
B	$0.89\Delta\omega$	$1.44\Delta\omega$	$1.3\Delta\omega$

表 5-1 中 A 为第一旁瓣衰减，B 为主瓣宽度(即峰值下降 3dB 时的带宽)，$\Delta\omega$ 为谱分析时角频率的分辨率。汉明窗由于其具有最大的第一旁瓣衰减和适中的主瓣宽度在实际音频处理中得到了广泛使用。

窗函数 $\omega(n)$ 的选择在很大程度上影响短时分析参数特性，因此为较好地反映音频信号特性，参数的选择和形状的确定应尽可能的合理。此外，窗函数长度 N 的选择尤为关键，窗口长度 N 如果很大，信号相当于通过很窄的低通滤波器，导致音频信号通过时高频部分被阻碍，短时能量随时间变化很小，不能真实地反映音频信号的幅度变化；反之，如果窗口长度 N 太小，滤波器的通带变宽，短时能量随时间有急剧的变化，不能得到平滑的能量函数。因此，窗口的长度选择应该合适，一般以 15～30ms 持续时间为宜。

5.2.4 时频分析

语音信号的时域分析就是分析和提取语音信号的时域参数。最常用的时域特征包括短时平均能量、

短时过零率、短时自相关函数和短时平均幅度差函数等，这是语音信号最基本的短时参数，在各种语音信号数字处理技术中都要应用，现分别讨论如下。

设经过加窗函数分帧处理后的音频数据的序列为 $x(n)$，将第 i 帧的音频数据记为 $x_i(n)$。此时短时平均能量 E_n 的计算公式如下：

$$E(i)=\sum_{n=0}^{N-1}x_i^2(n) \tag{5-5}$$

其中，N 表示第 i 帧的长度。短时平均能量可以反映音频数据随时间的能量分布。在此基础上，可以得到短时平均幅度 $M(i)$，其计算公式如下：

$$M(i)=\sum_{n=0}^{N-1}|x_i(n)| \tag{5-6}$$

短时平均幅度差 $D_i(k)$ 是由短时平均幅度拓展引出的，反映了相邻帧的幅度变化，可通过如下公式计算得到：

$$D_i(k)=\sum_{n=0}^{N-k-1}|x_i(n+k)-x_i(n)| \tag{5-7}$$

其中，k 为延迟量。

短时过零率 $Z(i)$ 表示一帧音频数据内幅值符号正负变化的次数，过零率反映了音频数据的波动水平，可用于背景噪声的检测，其计算公式如下：

$$Z(i)=\frac{1}{2}\sum_{n=0}^{N-1}|\operatorname{sgn}[x_i(n)]-\operatorname{sgn}[x_i(n-1)]| \tag{5-8}$$

其中，$\operatorname{sgn}[y_i(n)]$ 是符号函数。

短时自相关 $R_i(k)$ 主要反映了音频序列不同数据点之间的相关性，可用于端点检测，其计算公式如下：

$$R_i(k)=\sum_{n=0}^{N-k-1}x_i(n)x_i(n+k) \tag{5-9}$$

语音信号的频域分析就是分析语音信号的频域特征。下面介绍语音信号的傅里叶分析法。因为语音波是一个非平稳过程，因此适用于周期、瞬变或平稳随机信号的标准傅里叶变换不能用来直接表示语音信号。短时傅里叶变换可对语音信号的频谱进行分析，相应的频谱称为“短时谱”。利用短时傅里叶变换求语音的短时谱，其具体计算公式如下：

$$X_n(\mathrm{e}^{\mathrm{j}\omega})=\sum_{m=0}^{N-1}x(m)w(n-m)\mathrm{e}^{-\mathrm{j}\omega m} \tag{5-10}$$

其中，$x(m)$ 为音频序列，$w(n)$ 为窗序列函数，n 取不同的值时，窗 $w(n-m)$ 沿着时间轴滑动到相应位置，ω 为角频率，计算得到的不用音频帧的短时傅里叶变换可以反映音频信号的频谱随时间变化关系。

令 $X(\omega)$ 为 $x(n)$ 的傅里叶变换，则倒谱的计算公式如下：

$$c(n)=\mathrm{FT}^{-1}[\ln|X(\omega)|] \tag{5-11}$$

其中，$|X(\omega)|$ 为 $X(\omega)$ 的模值，FT^{-1} 为傅里叶逆变换。序列 $x(n)$ 经过对数幅值谱的傅里叶逆变换得到 $c(n)$，被称为倒频谱，简称倒谱。梅尔频率倒谱系数以人耳的听觉特性为基础，基于人耳听觉实验对音频频谱进行分析，从而提高音频特征的代表性。

5.3 音频数据表示

音频数据的表示是音频处理和分析的基础，不同的表示方法可以捕捉到音频信号的不同特征。本节将介绍两种主要的音频数据表示方法：离散表示法和分布表示法。离散表示法将音频数据转换为离散

的数字序列，以便计算机进行处理和存储。分布表示法则将音频数据表示为在时间和频率上的分布特征，以便更好地描述音频信号的时频特性。

5.3.1 离散表示法

离散表示法是音频分析中的一种基础技术，在处理音频信号时，通过一定的算法或变换提取出以离散数值形式表示的特征。音频特征提取是音频数据分布表示的第一步。常见的音频特征包括梅尔频率倒谱系数(MFCC)、频谱特征、节奏特征、音高特征等。这些特征能够反映音频的声学特性和感知特性，是后续分布表示的基础。

梅尔频率倒谱系数(MFCC)[4]的分析是基于人耳的听觉特性，主要依据人耳的听觉实验结果来分析音频的频谱，以获得具有更好代表性的音频特征。MFCC 特征能够有效地捕捉音频信号的重要特征，并且对噪声和语音变化具有一定的鲁棒性，通常用于语音识别、语音合成、说话人识别等任务。MFCC 分析所依据的人耳听觉特性有两个：

第一个特性是，人耳对频率的感知并不是线性的，下面的公式是人耳感知的频率和实际频率之间的关系：

$$F_{\text{mel}} = 2595\lg\left(1+\frac{f}{700}\right) \tag{5-12}$$

其中，F_{mel} 表示人耳感知的频率，f 表示实际频率，单位分别为梅尔(Mel)和赫兹(Hz)，梅尔和赫兹的关系如图 5-4 所示。

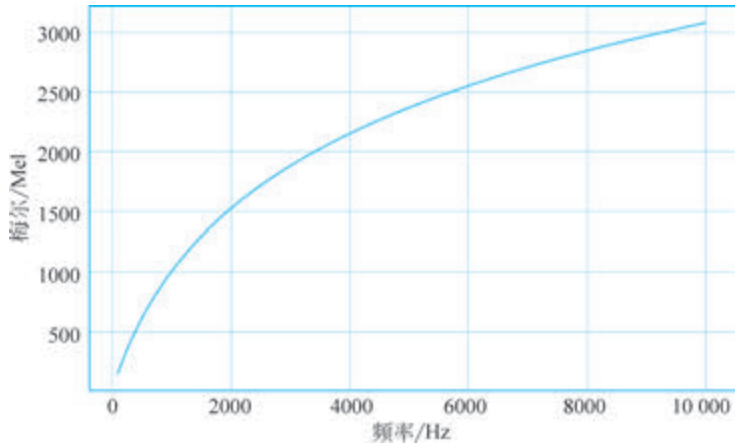

图 5-4 梅尔和赫兹的关系对应图

第二个特性是，相差较小的不同频率带被归为同一频率带，在大脑中会被叠加在一起识别处理，按这种频率群即临界带(critical band)划分，可以将音频在频域上分解为一系列的频率滤波器组，即梅尔滤波器组。计算 MFCC 的公式如下：

$$\text{MFCC}(i,n) = \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log\left[S(i,m)\right]\cos\left(\frac{\pi n(2m-1)}{2M}\right) \tag{5-13}$$

其中，$S(i,m)$表示梅尔滤波器能量，m 是第 m 个梅尔滤波器(共有 M 个)，i 表示第 i 帧，n 表示离散余弦变换(DCT)后的谱线。

如图 5-5 所示，MFCC 特征的计算通常包括以下步骤：

(1) 预处理。计算 MFCC 特征首先要对音频信号进行预处理，如前所述，包括音频清洗、分帧处理等。

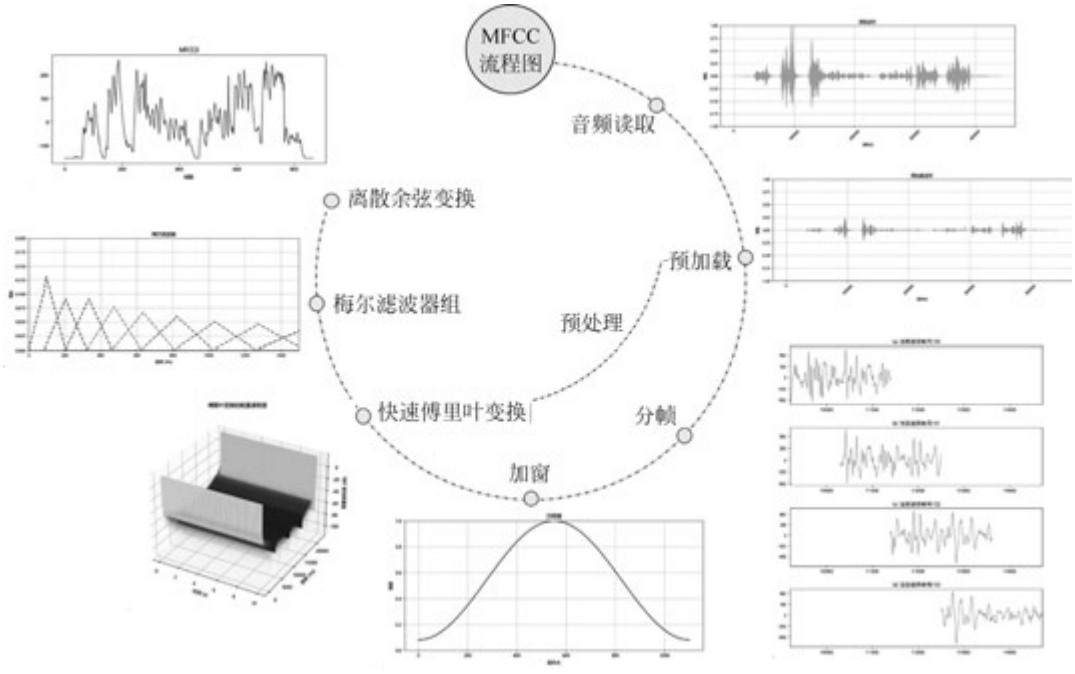

图 5-5 MFCC 计算过程

(2) 快速傅里叶变换(FFT)。对每个加窗后的帧进行 FFT,将信号从时域转换到频域,得到频谱的幅度谱或功率谱。FFT 的结果表示了信号在不同频率上的能量分布。

(3) 梅尔滤波器组。设计一组在梅尔刻度上均匀分布的三角形带通滤波器,这些滤波器在低频部分分布密集,在高频部分分布稀疏,以模拟人耳对频率的非线性感知特性。将 FFT 得到的幅度谱或功率谱通过这组滤波器,计算每个滤波器的输出能量。对每个滤波器的输出能量取对数,得到 log-Mel 滤波器组能量。这一步是为了进一步模拟人耳对声音的非线性感知,并使能量分布更加平滑。

(4) 离散余弦变换(DCT)。对 log-Mel 滤波器组能量进行 DCT,得到 MFCC 系数。DCT 将 log-Mel 滤波器组能量从梅尔频率域转换到倒谱域,得到的 MFCC 系数描述了音频信号在梅尔频率范围内的能量分布特性。

独立成分分析(Independent Component Analysis,ICA)是盲源信号分离的一种常用算法,其作用是将一个混杂了多种信号的信息分解为一个个独立的信号。提到独立成分分析就不得不说著名的“鸡尾酒会问题”。该问题描述的是一场鸡尾酒会中有 n 个人一起说话,同时有 n 个录音设备,问怎样根据 n 个录音文件恢复出 n 个人的原始语音。鸡尾酒会问题也称为盲源分离问题,ICA 就是针对该问题所提出的一个算法。

下面给出该问题的形式化描述。假设麦克风所收集的观测信号 $\boldsymbol{X}\in\mathbb{R}^{n\times m}$ 是源信号 $\boldsymbol{S}\in\mathbb{R}^{n\times m}$ 的线性混合:

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{S} \tag{5-14}$$

其中,m 是所采集的语音数据点个数。矩阵 $\boldsymbol{A}\in\mathbb{R}^{N\times N}$ 称为混淆矩阵。因此,通过找到混淆矩阵即可从观测信号中恢复源信号。

假设源数据 $\boldsymbol{S}_j$ 的概率分布为 $p_S(\boldsymbol{S}_j)$,各信号源之间是相互独立的,因此整体信号的联合概率分布为 $p(\boldsymbol{S})=\sum_{j=1}^{n}p_S(\boldsymbol{S}_j)$。考虑概率密度函数 $p_x(x)$ 与概率累计函数 $F_x(x)$ 的导数关系:

$$\begin{cases} F_x(x)=P(\boldsymbol{AS}\leqslant x)=P(\boldsymbol{S}\leqslant \boldsymbol{A}^{-1}\boldsymbol{X})=F_S(\boldsymbol{A}^{-1}\boldsymbol{X}) \\ p_x(\boldsymbol{x})=\dfrac{\mathrm{d}F_x(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}=p_x(\boldsymbol{WX})\mid \boldsymbol{W}\mid \end{cases} \tag{5-15}$$

基于此，可得 $p(\boldsymbol{S})=\sum_{j=1}^{n}p_S(\boldsymbol{S}_j)=\sum_{j=1}^{n}p_S(\boldsymbol{W}_j\boldsymbol{X})\mid \boldsymbol{W}\mid$。令概率累计函数为 Sigmoid 函数 $g(x)=\dfrac{1}{1+\mathrm{e}^x}$，则概率密度函数为 $p(x)=g'(x)=g(x)(1-g(x))$。基于此可定义数据似然为

$$L=\sum_{i=1}^{m}\sum_{j=1}^{n}p_x(\boldsymbol{W}_i\boldsymbol{X}^{(i)})\mid \boldsymbol{W}\mid \tag{5-16}$$

基于随机梯度上升的迭代优化，可得到满足最大似然的混淆矩阵 $\boldsymbol{W}$。

5.3.2 分布表示法

在分布表示中，每个音频片段或音频特征被映射到一个低维向量空间中，使得相似的音频在向量空间中具有相近的表示，不相似的音频则距离较远。这种表示方法不仅降低了音频数据的维度，还保留了音频的关键信息和特征，便于进行后续的处理和分析。深度神经网络模型，特别是卷积神经网络和循环神经网络(RNN)及其变体(如 LSTM、GRU 等)，在音频数据的分布表示中得到了广泛应用。这些模型能够自动从音频数据中学习特征表示，并通过非线性变换将特征映射到低维向量空间中。其中，CNN 擅长捕捉音频的局部特征，如频谱图中的纹理信息。RNN 的关键在于其循环结构，它允许网络在处理当前时间步的输入时，能够利用之前时间步的信息。这种特性使得 RNN 能够捕捉音频信号中的长期依赖关系，从而在处理如语音识别、音频分类等任务时表现出色。需要注意的是，传统的 RNN 结构在处理长序列时容易遇到梯度消失或梯度爆炸的问题，这限制了其在某些任务中的应用。为了解决这个问题，研究人员提出了多种改进结构，如 LSTM(长短期记忆网络)和 GRU(门控循环单元)等。

如图 5-6 所示，LSTM 网络通过引入门控机制(包括遗忘门、输入门和输出门)以及 Cell 状态来解决传统循环神经网络(RNN)中的梯度消失和梯度爆炸问题。LSTM 网络的输入通常是一系列经过预处理和特征提取的音频特征序列。这些特征序列能够反映音频信号的关键信息，如频率、能量、时长等。比如经过前文所述的音频数据预处理与离散表示法得到的 MFCC 特征。

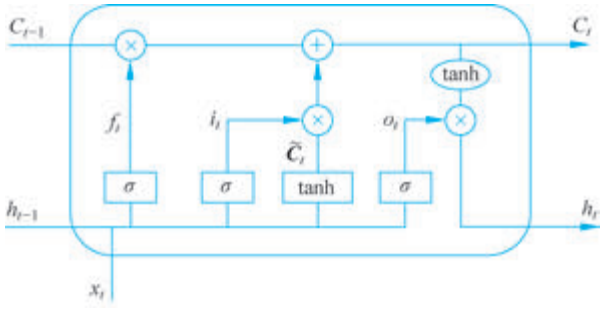

图 5-6 LSTM 网络结构

LSTM 模型在处理音频数据时，主要依赖于其独特的结构来捕捉音频中的时间依赖关系。LSTM 模型的基本结构包括遗忘门、输入门和输出门。

(1) 遗忘门。决定了上一时刻的 Cell 状态中有多少信息需要被遗忘。其公式为

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+b_f) \tag{5-17}$$

其中，f_t 是遗忘门的输出，σ 是 sigmoid 激活函数，$\boldsymbol{W}_f$ 是遗忘门的权重矩阵，b_f 是偏置项，h_{t-1} 是上一时刻的隐藏状态，x_t 是当前时刻的输入(在音频处理中，这通常是经过特征提取后的音频特征)。

（2）输入门。输入门决定了当前时刻的输入有多少需要被更新到 Cell 状态中。同时，Cell 候选状态计算了当前时刻的候选 Cell 状态值。其公式分别为

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t] + b_i) \tag{5-18}$$

$$\widetilde{C}_t = \tanh(\boldsymbol{W}_C \cdot [h_{t-1}, x_t] + b_C) \tag{5-19}$$

其中，i_t 是输入门的输出，$\widetilde{C}_t$ 是 Cell 候选状态，$\boldsymbol{W}_i$ 和 $\boldsymbol{W}_C$ 是相应的权重矩阵，b_i 和 b_C 是偏置项。新的 Cell 状态由遗忘门和输入门共同决定，其公式为

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{5-20}$$

其中，C_{t-1} 是新的 Cell 状态，C_{t-1} 是上一时刻的 Cell 状态，* 表示逐元素乘法。

（3）输出门。输出门决定了 Cell 状态中有多少信息需要被输出到隐藏状态中。其公式为

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, x_t] + b_o) \tag{5-21}$$

$$h_t = o_t * \tanh(C_t) \tag{5-22}$$

其中，o_t 是输出门的输出，h_t 是当前时刻的隐藏状态，$\boldsymbol{W}_o$ 是输出门的权重矩阵，b_o 是偏置项。

在构建 LSTM 模型时，通常会选择合适的网络结构和参数配置，以适应特定的音频处理任务。使用包含音频特征和对应标签的训练数据集对 LSTM 模型进行训练。在训练过程中，LSTM 模型会学习如何从输入的音频特征中提取出有用的隐含特征，并建立这些特征与输出标签之间的映射关系。通过训练好的 LSTM 模型，可以对新的音频信号进行特征提取和表示。在 LSTM 模型中，隐藏层的输出可以被视为音频信号的隐含特征表示。这些隐含特征不仅包含了音频信号的基本信息，还蕴含了时间序列之间的长距离依赖关系。

5.4 小结与延伸阅读

本章介绍了音频数据的基本概念，强调了音频数据的连续性和时序性特点，介绍了音频数据预处理的重要性，并概述了音频数据的表示方法。

近年来，基于 Transformer 架构的音频大模型在近年来得到了快速发展，这些模型利用 Transformer 的自注意力机制，有效地处理了音频数据中的复杂信息，实现了音频识别、生成、分类等多种任务。Whisper[9]是 OpenAI 推出的一个开源音频模型，Whisper 使用了 Transformer 架构，并结合了先进的音频处理技术，能够处理多种语言的语音数据，进行自动语音识别和语音到语音的翻译。Stable Audio Open 的架构由自动编码器、基于 T5 的文本嵌入以及基于 Transformer 的扩散模型（DiT）[10]构成。自动编码器将波形数据压缩到可管理的序列长度，T5 模型将输入的文本转换成文本嵌入，而 DiT 模型则在自动编码器的潜在空间中运行，对编码器压缩后的数据进行处理和优化。

习题

1. 音频信号的基本特性是什么？（　　）
   A. 它是离散的数字信号
   B. 它是连续的时域信号，具有波形、频率、振幅等特征
   C. 它只包含高频成分
   D. 它仅通过数字信号处理器生成
2. 在音频处理中，使用哪种技术可以去除或降低音频信号中的噪声？（　　）

A. 傅里叶变换　　B. 噪声滤波

C. 支持向量机(SVM)　　D. 逻辑回归

3. 简述傅里叶变换在音频处理中的应用。

4. 什么是音频预处理？列举几个常见的预处理步骤。

5. 什么是音频信号的时频表示？为什么它在音频处理中很重要？

参考文献

[1] 韩纪庆,张磊,郑铁然. 语音信号处理[M]. 3 版. 北京：清华大学出版社,2019.

[2] 胡航. 现代语音信号处理[M]. 北京：电子工业出版社,2014.

[3] Quatieri T F. 离散时间语音信号处理[M]. 赵胜辉,译. 北京：电子工业出版社,2004.

[4] 赵力. 语音信号处理[M]. 北京：机械工业出版社,2016.

[5] 赵一凡,卞良,丛昕. 数据清洗方法研究综述[J]. 软件导刊,2017,16(12)：222-224.

[6] 濮勇. 基于深度学习的音频场景分类[D]. 南京邮电大学,2021. DOI:10. 27251/d. cnki. gnjdc. 2021. 001164.

[7] 张力文. 时序关系依赖的音频语义特征学习方法研究[D]. 哈尔滨：哈尔滨工业大学,2021.

[8] Shi X,Chen Z,Wang H,et al. Convolutional LSTM Network：A Machine Learning Approach for Precipitation Nowcasting. NeuIPS,2015.

[9] Radford A,Kim J W,Xu T,et al. Robust speech recognition via large-scale weak supervision[C]//International Conference on Machine Learning. PMLR,2023：28492-28518.

[10] Peebles W,Xie S. Scalable diffusion models with transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023：4195-4205.