《语音特征提取》讲义

语音特征提取是语音数字信号处理中的一个核心环节，它涉及从语音信号中提取具有代表性和区分性的特征参数，以便用于语音识别、语音合成、说话人识别等应用。本节将学习常用的语音特征提取方法，包括短时能量、短时过零率、梅尔频率倒谱系数（MFCC）等。

# 一、语音特征提取概述

## 语音特征概念与分类

在语音信号处理中，语音特征是声音信号中具有特定频率、幅度和时域特征的数字表示，用于描述和区分不同语音信号。我们通常会从三个主要领域提取特征：时域、频域和倒谱域。这些特征帮助我们更深入地理解和分析语音信号，从而在各种应用中实现更准确的语音识别、说话人识别、以及声音质量分析等。

（1）时域特征

过零率 (Zero Crossing Rate, ZCR)：这是测量信号符号变化的速率。简单来说，它计算了信号在正负之间变化的频率，对于区分语音和非语音段非常有用。

短时能量 (Short-Time Energy)：这表示信号在一段时间内的能量总和。它是声音活动检测中的一个关键特征，可以帮助我们识别哪些部分是语音，哪些部分是静默。

短时自相关函数 (Short-Time Autocorrelation Function)：通过衡量信号与其时间延迟副本的相似度，这个特征对于基频的检测非常有用。它可以帮助我们识别语音信号中的周期性成分。

短时平均幅度差 (Short-Time Average Magnitude Difference)：这测量了信号幅度变化的平均值，用于分析声音的纹理特征，从而可以揭示声音的一些细微差别。

（2）频域特征

语谱图 (Spectrogram)：基于快速傅里叶变换（FFT），语谱图描述了信号随时间变化的频率和强度。它是可视化语音特征的一个强大工具，可以揭示语音的动态变化。

短时功率谱密度 (Short-Time Power Spectral Density)：这描述了信号在各个频率上的功率分布，对于分析声音的质量和特性非常重要。

频率倾斜 (Spectral Slope)：通过衡量频谱能量分布的倾斜度，这个特征用于描述声音的亮度或暗度，有助于我们理解声音的频率特性。

谱线频率：通过识别频谱中的显著频率成分，如基频和谐波，这些成分对于音高感知和音色分析至关重要。

谱熵 (Spectral Entropy)：衡量频谱能量分布的均匀性或随机性，通过计算频谱的概率分布的熵来得到，用于分析声音的复杂性和纹理。

（3）倒谱域特征

梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCCs)：通过对声音的短时功率谱进行对数非线性表示，MFCCs 提供了一个描述声音特性的强大工具，广泛用于语音和说话人识别。

线性预测倒谱系数 (Linear Predictive Cepstral Coefficients, LPCCs)：基于线性预测模型的倒谱表示，LPCCs 用于语音编码和识别，提供了一种有效的声音模型。

感知线性预测特征PLP (Perceptual Linear Prediction)：模拟人耳的听觉感知特性来提取语音特征，PLP技术考虑了听觉感知的复杂性，使得特征更贴近人类的听觉系统。

常数Q倒谱系数CQCC (Constant Q Cepstral Coefficients)：基于恒Q变换的倒谱系数，适用于处理具有变化频率分辨率的信号，特别适合于音乐和语音信号处理，提供了一种对信号进行更细致分析的方法。

通过这些特征的分析和应用，我们能够深入理解语音信号的复杂性，从而在各种场景下实现更加精准的声音识别和分析。

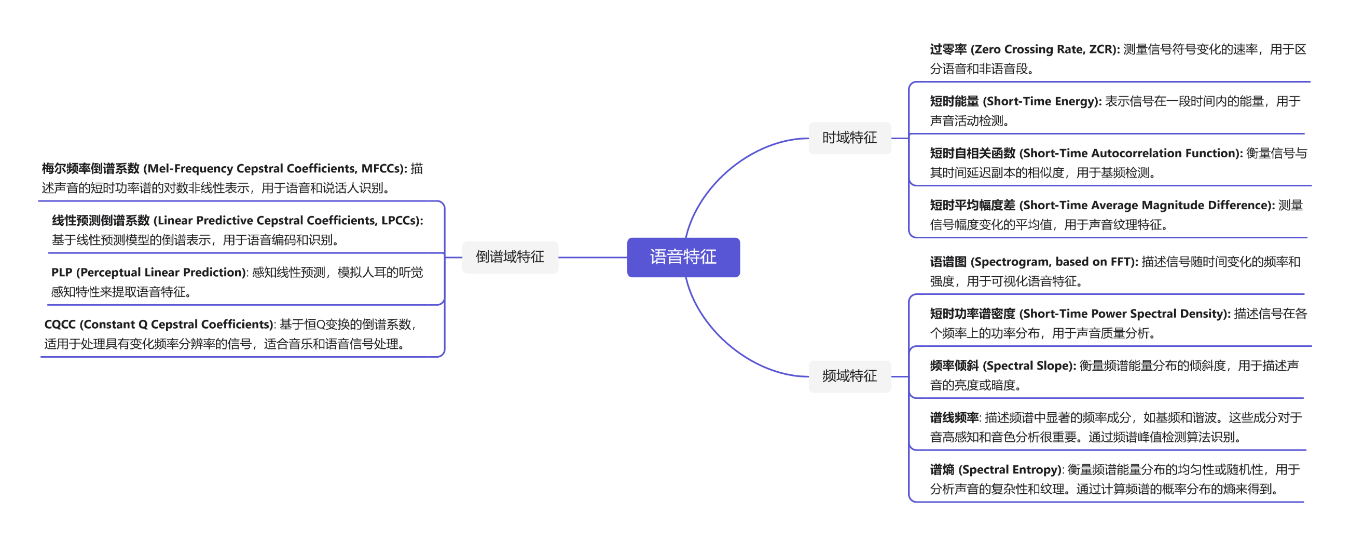


图1 语音的时域、频域及倒谱域特征概述

## 语音特征提取方法概述

以上三个域中，每个域都有其独特的方法和技术，下面我们就来一一解析这些方法。

首先是时域特征。这些特征是直接从语音信号的时间序列中提取的，非常直观且操作简单。想象一下，把语音信号切成一小段一小段的，然后用放大镜去观察每一小段的细节。通过将信号分割成短时帧（就是那些小时间窗口），可以对每一帧计算出一些有用的信息，比如信号的统计量、能量、动态变化和自相似性。这个方法不需要任何复杂的变换，直接在时域内完成所有操作。

接下来是频域特征，这里的操作就像是给语音信号换了一个“眼镜”，让我们能从频率的角度来观察信号。通过使用快速傅里叶变换（FFT）或其他相关变换，就可以将时域信号转换到频域中。一旦信号转换完成，就可以进行频率分析了，比如识别谐波频率、计算功率谱密度等。此外还会用到频率映射，比如将频率映射到梅尔刻度上，这是为了模拟人耳对不同频率的非线性感知。

最后探讨倒谱域特征的提取。这部分有点像是“魔法”，因为它通过对频域特征进行一系列变换，揭示了语音信号的一些深层次特性。首先是对数变换，它主要针对频域特征（如功率谱）进行，目的是强调信号的谐波结构和共振峰。然后，通过对这些对数变换后的信号进行逆傅里叶变换或离散余弦变换（DCT），可以得到倒谱表示。基于这个表示，可以进一步提取出对语音处理非常重要的特征，如梅尔频率倒谱系数（MFCCs）或线性预测倒谱系数（LPCCs）。

通过这些方法，我们能够从原始的语音信号中提取出有价值的特征，这些特征对于后续的语音识别、说话人识别、情感分析等任务至关重要。

# 二、语音特征提取方法

## 2.1 时域特征提取

### 2.1.1 短时能量提取

时域特征中首先计算短时能量。短时能量的计算方法是一种基于信号的时域特征提取方法，通过计算一段时间内的信号能量，来反映该时段内语音信号的强度和能量。该方法的核心是对信号进行窗函数加权，以反映不同时间点的信号贡献不同的特点。

如果想要计算*n*时刻的短时能量，可以基于时刻 *n* 之前的 *N* 个点，即时刻 *n* 的能量是计算从区间 *[n-N+1, n]* 的能量。在这个区间内，对于每个时间点 *m*（*m* 属于区间 *[n-N+1, n]*），短时能量计算公式为：



其中  表示信号在时刻 *m* 的幅值。窗函数  的作用是为信号中的每个点分配一个权重，使得在计算短时能量时，信号中不同点的贡献不同。 是窗函数的参数，用于计算在时刻 *n* 的短时能量时，时刻 *m* 的语音信号的权重。通常，离时刻 *n* 越近的点具有较大的权重，而离时刻 *n* 较远的点具有较小的权重。短时能量公式中的  表示在时刻 *m* 的语音信号经过窗函数加权后的能量，这个能量值会被累加到整个窗口的短时能量中。最终得到的短时能量值是在整个窗口内的信号能量的平均值，这个平均值会被归一化，即除以窗口大小 *N*，以保证短时能量的大小和窗口大小 *N* 无关。右图是由幅度谱变为能量谱的一个示例。有时在计算过程中，窗口的起始和结束位置可能会超出信号范围。为了解决这个问题，通常采用零填充（Zero Padding）的方法在信号的起始和结束位置添加零。

图片包含 图形用户界面

描述已自动生成

图2 能量谱的计算

### 2.1.2 短时过零率提取

接下来是短时过零率（Short-Time Zero-Crossing Rate，简称 STZCR）。它是一种衡量语音信号在特定时间范围内过零点数量的特征。过零点是指语音信号从正值变为负值，或从负值变为正值的点。其计算公式为：





短时过零率的计算原理主要包括以下几个方面：

（1）选取窗口：首先选取一个合适的窗口大小 *N*，窗口大小决定了计算短时过零率时要考虑的信号范围。通常情况下，窗口大小需要根据具体的应用场景和信号特点来选择。

（2）计算过零点：遍历窗口内的每个点，计算相邻两点的符号函数之差的绝对值。如果两个相邻点的符号不同（一个为正，一个为负），那么它们之间的符号函数之差的绝对值为 2，表示这两个点之间存在一个过零点；否则，符号函数之差的绝对值为 0，表示这两个点之间没有过零点。

（3）窗函数加权：在计算短时过零率时，可以引入窗函数 *w(n)* 来对信号的不同部分分配权重。通过使用窗函数，可以使得计算结果更加平滑，减少过零率波动，有利于后续的语音处理任务。常用的窗函数有矩形窗、汉明窗、汉宁窗等。

（4）求和：将计算得到的每个点的过零点信息相加，然后乘以 1/2，得到窗口内的过零点总数。在短时过零率的计算公式中，乘以 1/2 是为了得到正确的过零点数量。在计算过零点时，我们比较相邻两点的符号函数之差的绝对值。当两个相邻点的符号不同（一个为正，一个为负），它们之间的符号函数之差的绝对值为 2，表示这两个点之间存在一个过零点。这里的“2”实际上表示有两个方向的过零点（从正到负，从负到正），但我们只关心单个过零点的数量。因此，我们需要将这个值除以 2，将其转换为实际的过零点数量。所以，在短时过零率的计算公式中，我们乘以 1/2 是为了将相邻点之间的符号函数之差的绝对值（可能的值为 0 或 2）转换为实际的过零点数量（可能的值为 0 或 1）。这样，我们就得到了在时刻 n 的短时过零率。

（5）重复计算：将窗口沿着语音信号滑动，重复上述过程，计算整个信号的短时过零率。

短时过零率这个特征在语音信号处理领域具有广泛的应用，如语音端点检测、语音分割、声音分类等。

图片包含 图形用户界面

描述已自动生成

图3 短时过零率的计算

### 2.1.3 短时自相关系数提取

最后是短时自相关函数（Short-Time Auto-correlation Function, STACF）。它帮助我们量化语音信号在特定时间范围内与自身的相似度，这种技术在很多应用中都非常有用，比如在检测语音信号的基音周期（也就是Pitch period）或者在语音编码方面。那么，短时自相关函数是怎么计算的呢？首先，我们用一个公式来表示它：



在这个公式中，表示在时刻的短时自相关函数的值，是延迟值，是在时刻的幅值，而则是窗函数的值，用于在计算时加权。要计算这个函数，我们需要按照几个步骤来操作：

1. 选取窗口：首先，我们需要选取一个合适的窗口大小。窗口大小非常关键，因为它决定了在计算短时自相关函数时要考虑的信号范围。选得太小，可能捕捉不到足够的信息；选得太大，又可能引入不必要的计算复杂度。

2. 计算自相关：接下来，对于每一个延迟值，我们计算窗口内信号值与延迟个点的信号值的乘积，然后把所有这些乘积求和。通过这种方式，我们就能得到在时刻的短时自相关函数的值。这个过程本质上是在测量信号在当前时刻与它自己在过去某个时刻的相似度。

3. 重复计算：为了得到整个语音信号的自相关特征，我们将这个窗口沿着语音信号滑动，重复上述计算过程。通过这种方式，我们可以得到信号在不同时间点的自相似性信息，这对于后续的分析非常有帮助。

短时自相关函数的美妙之处在于它简单而有效，能够揭示出语音信号中的周期性特征，比如基音频率，这对于很多语音处理任务来说是非常宝贵的信息。

## 2.2 频域特征提取

采集的语音都是时域上的，要提取频域特征首先要将时域的信号转化到频域上来，这一般要借助于FFT。FFT（快速傅里叶变换）是一种高效的算法，用于将时域信号转换为频域信号。FFT算法可以快速地计算出信号在频域上的能量分布，让我们能够更加直观地观察信号的特性。在语音信号处理中，FFT通常被用于计算信号的短时傅立叶变换。对于一段语音信号，我们通常将其分成若干个短时帧，每个帧的长度通常为10-30ms。然后对每个帧进行加窗，以减小边界效应，并计算每个帧的FFT，得到该帧的频谱信息。这样就得到了语音信号的短时傅立叶变换。

在短时傅立叶变换中，每个帧的频谱信息都可以表示为一个*K*维向量，其中*K*表示FFT的长度。每个元素表示信号在对应频率上的能量分布。通过分析不同频率上的能量分布，可以提取出信号的一些重要特征，如共振峰、语调等信息。例如，对于一位发火的人来说，他的语音信号可能会在高频范围内具有更高的能量分布，因为情感激动会导致声带振动更加剧烈，从而产生更高频率的声波。通过观察语音信号的频谱信息，我们可以更加准确地识别出这些特征，从而实现语音信号的分析、识别、增强等任务。



图4 FFT变换过程

### 2.2.1 语谱图提取

由FFT直接得到的最基本的频域特征就是语谱图。语谱图（spectrogram）是一种将语音信号在时间和频率上的表示方法。它将音频信号分成一系列短时窗口，每个窗口内的信号进行傅里叶变换得到频域能量分布，再通过一定的处理方式，将这些频域能量分布转换为对应时频区域内的颜色强度，最终组合成语谱图。在语谱图中，纵向为微小时间间隔内声波的频率分布，横向为各频率随时间的变化。通常采用颜色或灰度来表示频域能量的大小，颜色越深或灰度越大表示能量越强。语谱图可以直观地反映语音信号在时间和频率上的特征，如音调、共振峰等。例如，在一个nineteenth century的发音的语谱图中，可以看到低频区域分布更多更密集的能量，这是因为这段语音是男声发音，男声的声音在低频区域更加明显。语谱图在语音信号处理中有着广泛的应用。例如，语音增强、语音识别、说话人识别等任务中，语谱图可以用于提取信号的特征，从而实现相关的处理和分析。此外，语谱图还可以用于语音信号的可视化，方便观察语音信号在时间和频率上的变化。

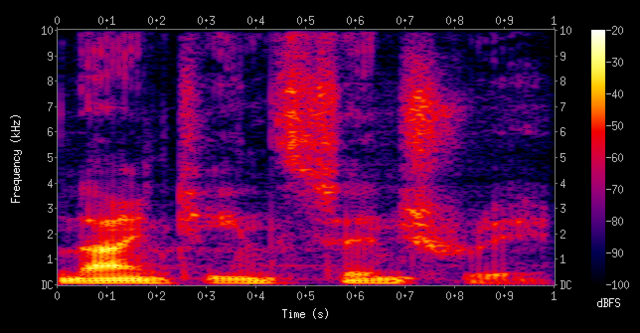


图5 nineteenth century发音语谱图

### 2.2.2 功率谱提取

接着继续加工就可以得到功率谱（Power Spectrum）。功率谱是指信号在频域上各个频率上的能量大小，通常使用傅里叶变换将时域信号转换到频域得到，然后对傅里叶变换结果的模值（因为频率不可能为负，负值要舍去）求平方即可得到功率谱。功率谱密度（Power Spectrum Density）是指单位频率范围内的信号功率，通常是对功率谱进行平滑化处理得到。在数字信号处理中，功率谱密度通常通过对信号进行窗口分析后，将窗口内的信号平方求和，然后除以窗口长度得到。具体地，对于一个长度为*N*的信号*x(n)*，做*N*点傅里叶变换得到的频域信号为*X(k)*，则其功率谱为*|X(k)|^2*，其中*k*表示频域上的索引。如果需要计算功率谱密度，则可以将所有的*|X(k)|^2*值按照频率分段进行平均，得到单位频率范围内的信号功率谱密度。功率谱和功率谱密度在语音信号处理中有着广泛的应用。例如，它们可以用于后续对语音信号的频率特性进行分析和提取，也可以用于语音信号的增强和去噪等任务。

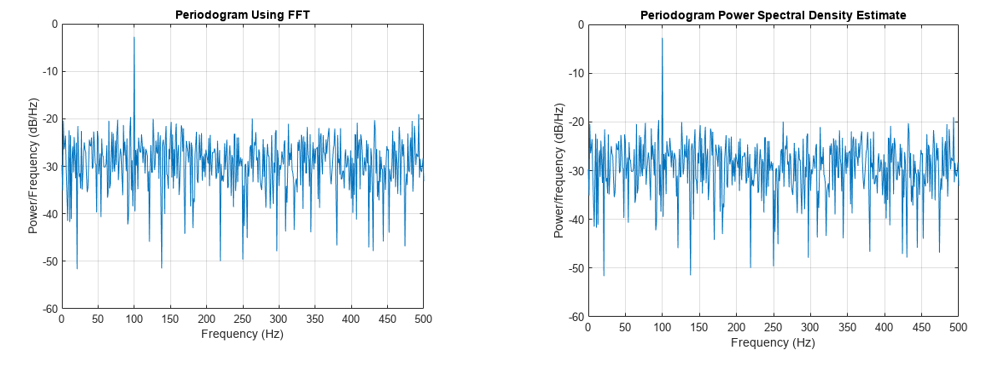


图6 功率和功率谱

### 2.2.3 其他特征

除了这些基本的频域特征外，还有一些其他用于不同语音应用的特征，在这里做个简介，不做详细了解，实际他们是对基本频谱的一些简单运算，体现某种特性。

1. 谱线频率：谱线频率是指频谱中的显著频率成分，它们代表了信号的基频和谐波。这些频率成分对于识别和分析音频信号的基本频率及其谐波成分至关重要。我们通常通过峰值检测算法从频谱中识别出显著的频率峰值来找到它们。

2. 频率倾斜：频率倾斜度量了频谱中高频成分与低频成分的能量比。它反映了声音的“亮度”或“暗度”，从而影响音色的感知。计算频率倾斜的方法是通过对频谱进行线性回归分析，然后计算频谱的线性拟合斜率。

3. 谱质心：谱质心是频谱能量的“重心”，可以理解为频谱的平均频率。它描述了声音的亮度或尖锐度，对音色的感知有直接影响。谱质心的计算方法是：



其中是频率，是该频率的幅度。

4. 谱扁平度：谱扁平度用于度量频谱的平坦程度，即所有频率成分的能量分布的均匀性。它有助于区分噪声和音调信号，评估声音的纹理。计算公式为：



这里也是幅度。

5. 谱熵：谱熵描述了频谱能量分布的随机性或均匀性，可以用来分析声音的复杂性和不确定性，特别是在声音纹理分析中非常有用。计算方法是：



其中是归一化能量，确保所有的和为1。

## 2.3 倒谱域特征提取

### 2.3.1 Fbank提取

在我们深入倒谱域特征的世界之前，让我们先搞清楚一件事：在数字语音处理中，这些特征不仅仅是数据点，它们是我们理解和解码人类语音的关键。倒谱域中我们先介绍一种较为简单的特征提取方法FBank（Filterbank）。这是一种常用的语音信号特征提取方法，它是基于梅尔滤波器的一种声谱分析方法。基于此得到的语音特征就是FBank特征。这里面的核心概念是梅尔滤波器。在介绍梅尔滤波器前，先介绍一下梅尔刻度。梅尔刻度（Mel Scale）是一种基于人耳听感知的刻度，能够更好地模拟人耳的频率分辨率。梅尔刻度与线性的频率刻度赫兹（Hz）之间可以进行近似的数学换算。

在梅尔刻度中，将1000 Hz，且高于人耳听阈值40分贝的声音信号，定为1000 mel的参考点。在频率500 Hz以上时，随着频率的增加，人耳每感觉到等量的音高变化，所需要的频率变化愈来愈大。这导致在赫兹刻度500 Hz往上的四个八度（一个八度即为两倍的频率），只对应梅尔刻度上的两个八度。梅尔刻度可以用于语音信号处理中的一些任务，如语音识别、说话人识别、情感识别等。它可以更好地模拟人耳对于音高的感知，提高信号处理的效果。将梅尔刻度转换为赫兹刻度的公式为：



其中，f\_Hz表示赫兹刻度上的频率，m表示梅尔刻度上的频率。将赫兹刻度转换为梅尔刻度的公式为：



其中，f\_Hz表示赫兹刻度上的频率，m表示梅尔刻度上的频率。梅尔频谱图如下：

图表

描述已自动生成

图7 梅尔刻度

Mel滤波器组是在Mel刻度下均匀分布的一组滤波器，通常用于将语音信号分解成Mel频谱。Mel滤波器组的数量和中心频率位置是预先确定的，通常使用对数刻度而不是线性刻度，以模拟人耳的感知。如图是一个有M个三角滤波器的Mel滤波器组（滤波器的个数和临界带的个数相近），M通常取22-40，26是标准。每个滤波器都是三角形的，中心频率为f(m)，其中m是Mel刻度上的频率，f(m)可以通过将Mel刻度上的频率转换为赫兹刻度上的频率后得到。中心频率处的响应为1，并向0线性减小，直到达到两个相邻滤波器的中心频率，其中响应为0，各f(m)之间的间隔随着m值的增大而增宽。

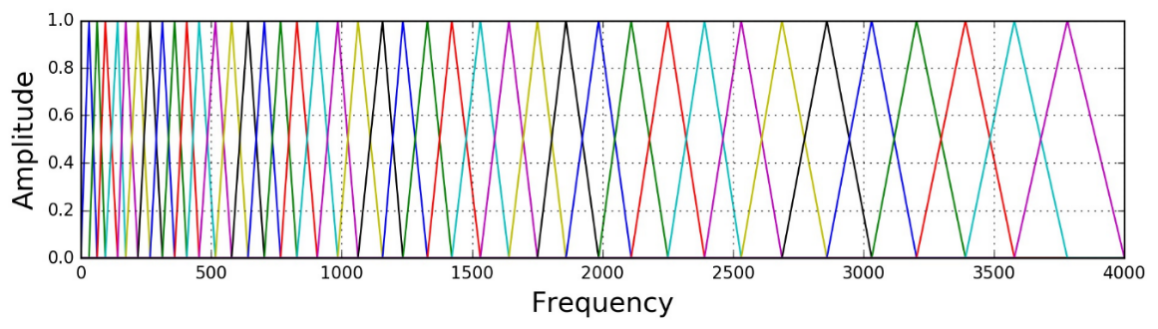


图8 梅尔滤波器组

Mel滤波器组的作用是将语音信号分解成一组Mel频谱系数，用于语音信号的分析和处理。它可以模拟人耳的感知，将语音信号在Mel刻度下的频率特性进行描述，提取出更符合人耳听觉特性的特征，提高语音信号处理的效果。在语音识别、说话人识别、情感识别等任务中，Mel滤波器组通常作为特征提取的一部分，用于提取出信号中的重要信息，进而实现相关的处理和分析。

接下来介绍FBank特征提取过程。关键步骤可以概括为以下几个步骤：

（1）对语音信号进行短时傅里叶变换，得到每个时间窗口内信号的频谱。

（2）对每个时间窗口的频谱进行幅度平方运算，即可得到该时间窗口内每个频率分量的功率谱。

（3）将功率谱除以信号采样周期得到的功率谱密度，更好表示了每个频率分量的能量分布情况。

（4）将功率谱密度与梅尔滤波器组进行卷积运算，即可得到每个时间窗口内每个滤波器的输出能量。

（5）将所有滤波器输出能量求和，即得到每个时间窗口内音频信号的总能量。

（6）对于每个时间窗口，取其总能量的对数值作为FBank系数。

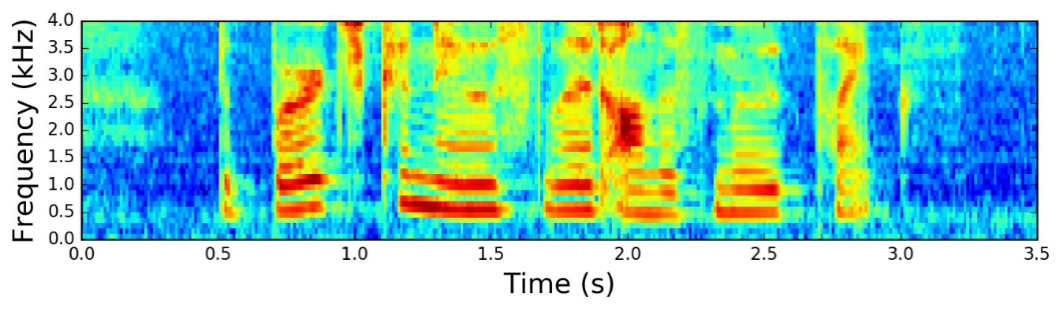


图9 FBank特征（系数）

放到整个语音处理的过程中来看，FBank的整体提取流程为：

（1）预处理：对原始语音信号进行预处理，例如进行加重处理，以增强高频信号的能量。

（2）分帧加窗：将预处理后的语音信号分成若干帧，每帧包含N个采样点，相邻帧之间有重叠部分。对每一帧进行窗函数加窗，以减少频谱泄漏。

（3）傅里叶变换：对每一帧信号进行短时傅里叶变换（Short-time Fourier Transform, STFT），得到每个时间窗口内信号的频谱（幅度谱）。

（4）梅尔滤波器组：将每帧的幅度谱与梅尔滤波器组进行卷积运算，得到每个时间窗口内每个滤波器的输出能量。

（5）对数运算：对每个时间窗口的所有滤波器输出能量求和，得到该时间窗口内音频信号的总能量。对总能量进行对数运算，得到FBank系数。（对数幅度谱或对数功率谱能够将信号的能量范围缩小到一个较小的区间内，使得能量变化较小的信号细节能够更好地被表示）

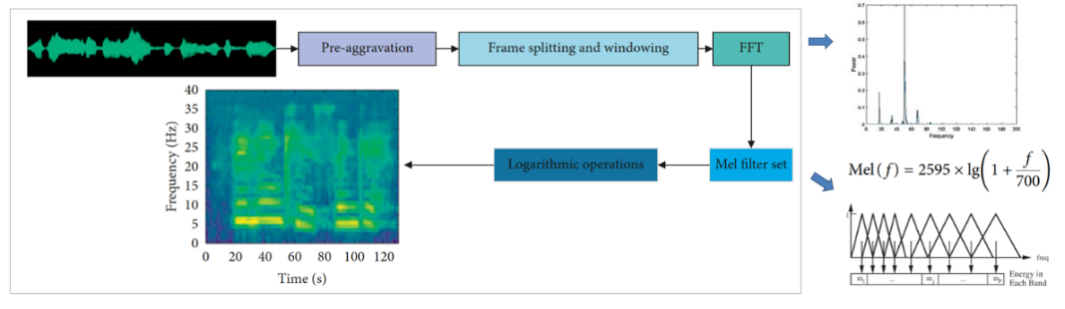


图10 Fbank提取流程图

### 2.3.2 MFCC提取

接下来的部分我们重点讨论一个在语音识别和处理领域极为重要的特征——梅尔频率倒谱系数MFCC（Mel-Frequency Cepstral Coefficients）。MFCC是基于Fbank（滤波器组）特征提取出来的。这个过程涉及到一系列步骤，从预处理开始，到分帧加窗，再到傅里叶变换，我们得到了频谱。紧接着，我们通过梅尔滤波器组过滤这个频谱，进行对数运算，最后通过离散余弦变换（DCT）来得到MFCC。这个过程听起来可能有点复杂，但其实它的核心目的非常直接：减少特征之间的相关性，并模拟人耳的非线性感知特性。

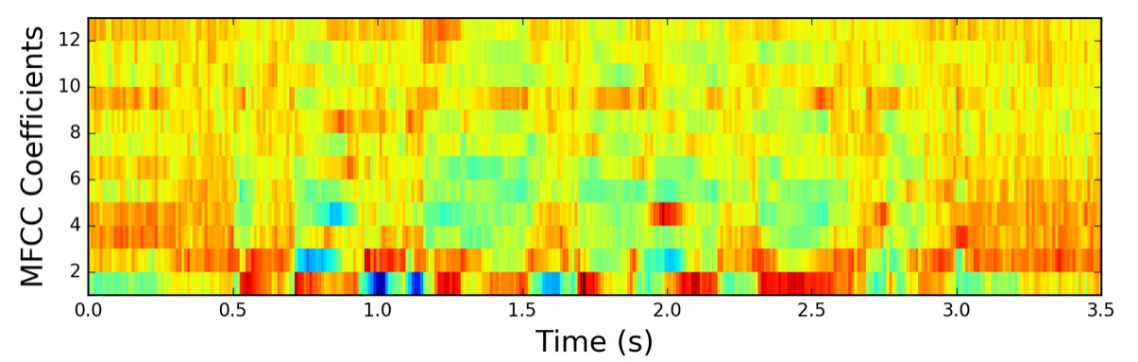


图11 MFCC系数

MFCC特征提取的主要步骤如下：

（1）预处理：对原始语音信号进行预处理，例如进行加重处理，以增强高频信号的能量。

（2）分帧加窗：将预处理后的语音信号分成若干帧，每帧包含N个采样点，相邻帧之间有重叠部分。对每一帧进行窗函数加窗，以减少频谱泄漏。

（3）傅里叶变换：对每一帧信号进行短时傅里叶变换（Short-time Fourier Transform, STFT），得到每个时间窗口内信号的频谱（幅度谱）。

（4）梅尔滤波器组：将每帧的幅度谱与梅尔滤波器组进行卷积运算，得到每个时间窗口内每个滤波器的输出能量。

（5）对数运算：对每个时间窗口的所有滤波器输出能量求对数，得到MFCC系数。

（6）DCT变换：对每帧MFCC系数进行离散余弦变换（Discrete Cosine Transform, DCT），得到MFCC特征向量。

（7）动态特征：对MFCC特征向量进行一些动态特征处理，例如一阶差分、二阶差分等，以提高特征的区分度和稳定性。

MFCC特征提取中的梅尔滤波器组和DCT变换是其核心部分。梅尔滤波器组能够提取出信号在梅尔刻度下的能量分布特征，更好地模拟人耳的听觉感知特性，从而提高了语音信号处理的效果。DCT变换能够将MFCC系数从时域转换到频域，提取出信号的谐波特征，进而提高特征的区分度和稳定性。离散余弦变换在这里扮演了一个英雄的角色。它帮助我们去除了信号中不同维度间的相关性，并将这些信号映射到了多维空间中。这就是为什么我们可以通过MFCC捕捉到语音信号的本质，因为它通过DCT消除了频率分量之间的高度相关性。

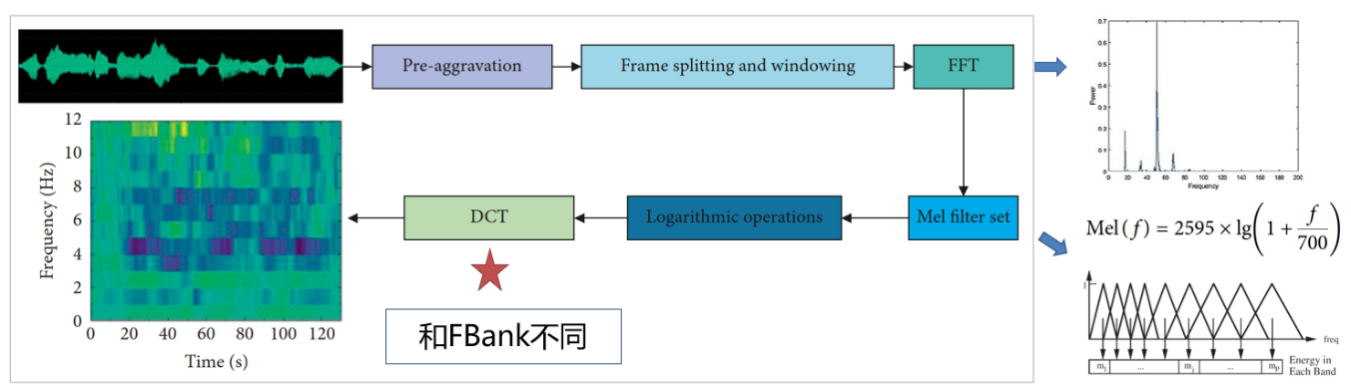


图12 MFCC提取流程图

MFCC和FBank特征是在语音处理和语音识别中常用的特征，它们具有不同的特点和应用场景。

（1）计算量：MFCC是在FBank的基础上进行的，所以MFCC的计算量更大。因为MFCC还需要进行DCT变换等额外的处理步骤。

（2）特征区分度：FBank特征相关性较高，相邻滤波器组有重叠，因此其特征区分度较低。而MFCC具有更好的判别度，因此在大多数语音识别论文中用的是MFCC，而不是FBank。

（3）信息量：FBank特征的提取更多的是希望符合声音信号的本质，拟合人耳接收的特性。MFCC做了DCT去相关处理，因此Filter Banks包含比MFCC更多的信息。

（4）模型适用：在GMM-HMM框架下，通常会采用对角协方差矩阵对GMM进行建模，这也就要求特征空间的每一维相互独立，而DCT变换的解相关特性可以很好的满足GMM的独立性假设。在深度学习模型中，DNN/CNN可以更好的利用Filter Banks特征的相关性，降低损失。

### 2.3.3 LPCC提取

接下来介绍几种应用也比较多的特性，首先是线性预测倒谱系数（Linear Predictive Cepstral Coefficients，简称LPCC）。这是一种基于线性预测编码（LPC）的声学特征提取方法，它在语音信号处理领域有着广泛的应用。LPCC结合了线性预测分析和倒谱分析，旨在提取出能够代表语音信号特性的参数。LPC的核心思想是假设一个信号可以通过其过去的样本来线性预测。在语音信号处理中，这通常意味着声道的当前输出可以通过其过去的输出来预测。LPC通过求解一个线性方程组来找到一组系数，这些系数描述了信号与其自身过去样本之间的关系。数学模型为：







以上第一个公式表示当前样本  可以通过过去  个样本的加权和来近似。这里， 是LPC系数， 是过去的样本。第二个公式的含义是预测误差  是实际样本  与通过LPC模型预测的样本之间的差值。这个误差代表了模型未能解释的信号部分。最后一个公式描述了如何递归地计算LPC系数 。对于 ， 就是  的对数。对于 ， 是  加上  个之前的系数  与  样本的加权和。对于 ， 是  个之前的系数  与  样本的加权和。

LPC的步骤通常包括：

1.自相关序列计算：计算信号的自相关序列，这是信号与其自身在不同时间延迟下的内积。

2.LPC系数求解：通过自相关序列求解LPC系数。这通常涉及到求解一个Yule-Walker方程组，或者使用更高效的Levinson-Durbin算法。

3.预测误差信号：使用LPC系数来计算预测误差信号，即实际信号与预测信号之间的差值。

而倒谱分析的步骤根据以上描述则主要括：

1.离散余弦变换（DCT）：对自相关序列进行离散余弦变换，得到倒谱系数。

2.倒谱系数：倒谱系数是自相关序列的DCT结果，它们可以用来重构信号的功率谱。

两者结合就是LPCC：

1. LPC系数求解：首先，对语音信号进行LPC分析，得到一组LPC系数。

2. 自相关序列：使用LPC系数计算自相关序列。

3. 倒谱分析：对自相关序列进行倒谱分析，得到LPCC。

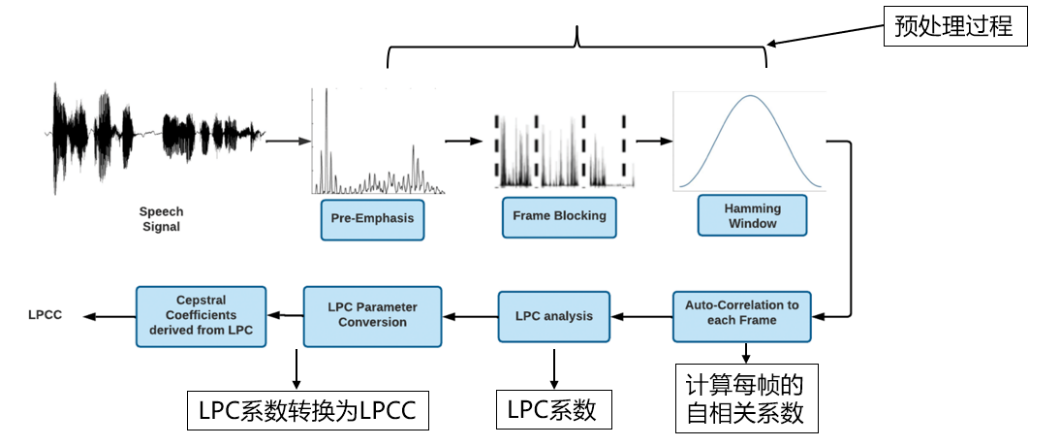


图13 LPCC提取流程图

LPCC提供了一种在时域和频域之间转换的方法，使得语音信号的特征更加稳定和鲁棒。这些特征对于语音识别、语音压缩、语音合成等应用至关重要，因为它们能够捕捉到语音信号的重要声学特性，如共振峰（Formants）和声道的谐波结构。LPCC的一个主要优势是它的计算效率。由于LPC和倒谱分析都是基于自相关序列的，所以LPCC可以在不进行完整的傅里叶变换的情况下提取特征，这在计算资源有限的情况下尤其有用。此外，LPCC特征对于语音信号的时变特性（如音调和语速变化）具有很好的鲁棒性。

### 2.3.4 PLP提取

PLP（Perceptual Linear Predictive）是一种模拟人耳听觉特性的语音分析方法，它通过一系列步骤来提取语音信号的特征，这些特征更接近于人类听觉系统对声音的处理方式。PLP的步骤主要包括：

1. 预加重：预处理步骤通常包括预加重，这是一种频率补偿过程，用于增强语音信号的高频部分。这是因为人耳对高频声音的灵敏度较低，预加重可以模拟人耳对高频的非线性放大效应。

2. 短时傅立叶变换（STFT）：对预处理后的语音信号进行短时傅立叶变换，以获得信号的频谱。这一步将信号从时域转换到频域，便于后续的频谱分析。

3. Bark滤波器组：使用Bark滤波器组对STFT得到的频谱进行处理。Bark滤波器组模拟了人耳的临界频带特性，将频谱映射到Bark刻度上。这有助于捕捉人耳对不同频率的敏感度。

4. 等响度预加重：根据等响度曲线对Bark滤波后的频谱进行预加重。等响度曲线反映了不同频率声音的响度感知，这一步进一步模拟了人耳对声音强度的非线性响应。这一步骤使用的是等响度曲线（Equal-Loudness Contours），通常参考的是国际标准ISO 226:2003中的曲线。等响度曲线是基于听觉实验得出的，描述了在不同频率下，声音需要达到的强度水平才能被人耳感知为相同的响度。具体目的就是调整频谱以反映人耳对不同频率声音的敏感度。在低频和非常高频区域，人耳对声音的感知不如中频区域敏感。通过应用等响度曲线，可以对频谱中的不同部分进行加权，使处理后的信号更接近人类听觉系统对声音的实际感知，可以认为就是一个设置好参数的调节器。

5. 强度-响度转换：对预加重后的频谱进行强度-响度转换，以模拟声音强度与听觉响度之间的关系。这通常涉及到对频谱能量进行非线性变换，如取立方根。在这一步骤中，对频谱能量进行非线性转换，以模拟声音强度与听觉响度之间的关系。人类的听觉感知不是线性的，意味着声音强度的增加不会导致响度感知成比例增加。通过应用强度-响度转换，可以确保提取的特征更贴近于人耳对响度的实际感知。

6. 线性预测分析（LPC）：使用LPC模型来估计声道的冲激响应。LPC是一种参数化方法，它通过最小化预测误差来确定一组系数，这些系数可以描述声道的共振特性（即共振峰）。

7. 倒谱分析：对LPC分析得到的预测误差信号进行倒谱分析，得到倒谱系数。倒谱系数是声道特性的一种表示，可以用于后续的语音识别或处理任务。

8. 离散余弦变换（DCT）：对倒谱系数进行离散余弦变换，以获得PLP特征。DCT有助于降低特征的维度，同时保留重要的声道信息。

每个步骤的作用都是为了更好地模拟人耳对声音的处理，提取出对语音识别和理解更有用的特征。PLP特征因其对噪声的鲁棒性和对人类听觉特性的良好模拟而在语音处理领域得到了广泛应用。

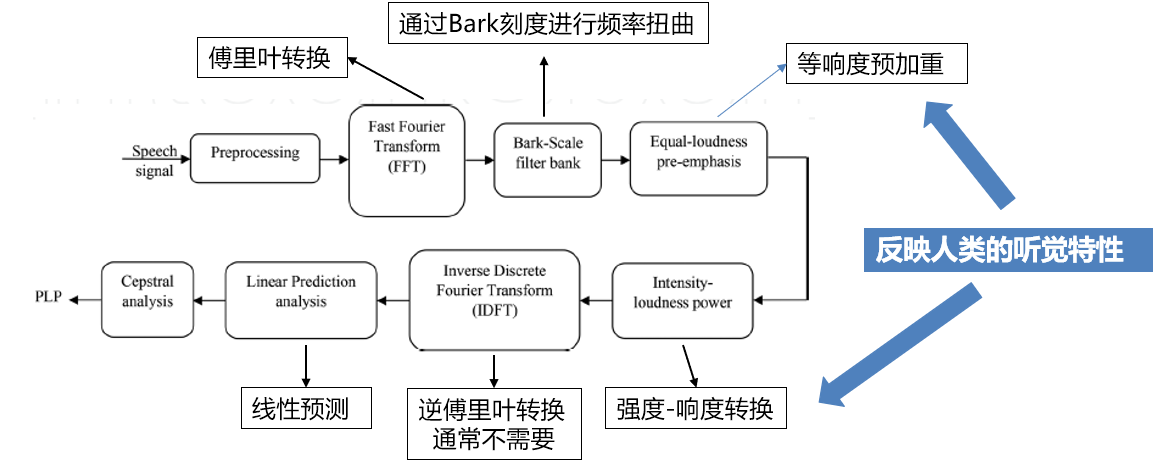


图14 PLP特征提取流程图

最后重点介绍一下Bark滤波器。Bark滤波器是基于Bark尺度设计的一组滤波器，用于模拟人耳对声音频率的非线性感知。Bark尺度是由德国声学家Eberhard Zwicker提出的，它将线性频率映射到心理声学的尺度上，这个尺度反映了人类听觉系统对不同频率声音的敏感度。Bark滤波器的设计通常涉及到将频率从Hz转换到Bark尺度，然后根据Bark尺度来确定滤波器的中心频率和带宽。转换公式如下：



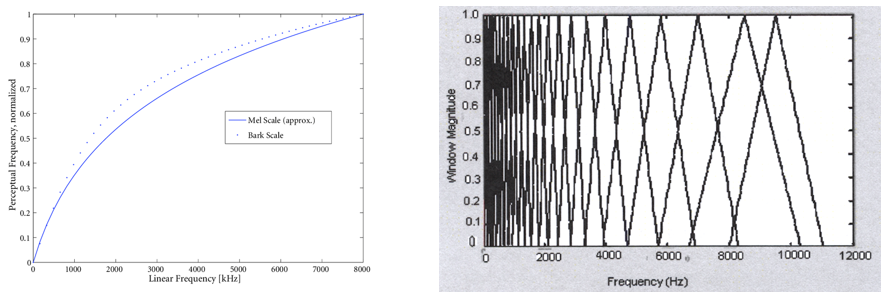


图15 Bark刻度与滤波器组

当然后续一些研究者也对Bark转换公式提出了不同的版本，在这里不再详述。我们主要解释一下Bark为何设计成以上那样奇怪的形式。在心理声学领域有一个概念：临界频带。人耳的临界频带特性是指在听觉系统中，当两个声音的频率相差一定范围时，它们会被感知为不同的音调。这个范围称为临界带宽，它是一个频率区间，在这个区间内，声音的能量可以被整合并感知为一个整体。临界带宽的大小随着中心频率的增加而增加，这意味着实际低频区域，人耳会更敏感地捕捉声音的不同之处；在高频区域，人耳对频率的分辨能力降低。Bark尺度是基于临界频带的概念设计的。它通过两个arctan函数的组合实现了非线性映射。第一个arctan项主要影响低频部分，而第二个arctan的平方项对高频部分的影响更大。如果公式难以理解，可以看最终的滤波器图，将语音将频率非线性地映射到Bark刻度，这种映射考虑了人耳的临界频带特性。它在低频区更加密集，在高频区则更加稀疏。这意味着在低频区，即使是小的频率变化也会在Bark尺度上产生较大的差异，而在高频区，只有较大的频率变化才会导致Bark值的显著变化。很多滤波器都为了模拟人耳的这种感知特性而产生，比如我们最后要讲的CQCC。

### 2.3.5 CQCC提取

简单说，常数Q倒谱系数CQCC结合了MFCC和PLP的优点。CQCC是为了解决MFCC在某些情况下可能无法充分捕捉语音信号的频谱包络特性而提出的。MFCC在语音处理领域非常流行，但它在处理非平稳信号时可能不够鲁棒。PLP则试图通过模拟人耳的听觉特性来改进MFCC，但PLP的计算相对复杂。CQCC试图在保持MFCC的简单性的同时，引入PLP的一些优点，以提高特征的鲁棒性和准确性。CQCC是这么处理特征的：

1. 预加重：增强信号的高频部分，以补偿人耳对高频的非线性衰减。

2. 窗函数：对信号进行分帧处理，并在每一帧上应用窗函数，以减少帧边界的不连续性。这两步都是常规步骤。

3. Q变换：使用常数Q变换将信号从时域转换到频域。这种变换在Mel频率尺度上提供均匀的频率分辨率，更好地模拟人耳的听觉特性。

4.功率谱计算：对每一帧的频谱进行平方，得到功率谱。这一步有助于压缩幅度值的范围，提高特征的鲁棒性。

5.均匀采样：在频域上对功率谱进行均匀采样，以便将频率域划分为相等的频带。这有助于在不同频率区域中保持一致的分析粒度。

6.离散余弦变换（DCT）：对均匀采样后的功率谱进行离散余弦变换，得到CQCC。这一步类似于MFCC中的操作，通过DCT将频域特征转换为倒谱特征，以减少特征维度和去除冗余信息。

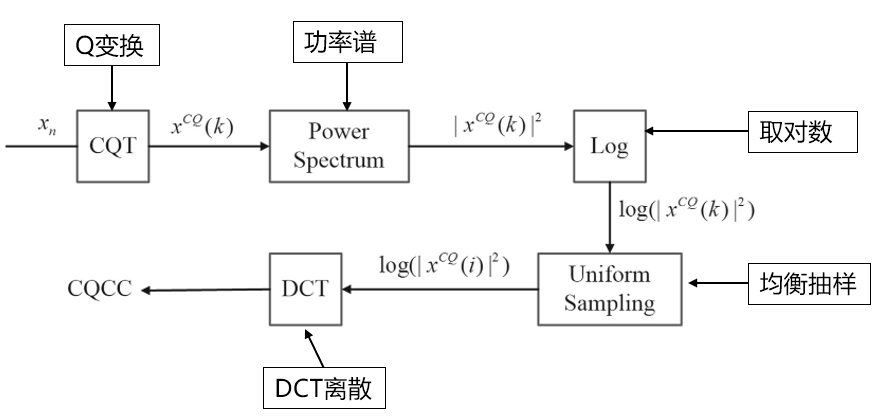


图16 常数Q变换特征提取

因此观察整个步骤，CQCC与MFCC区别主要在于两个部分：常数Q变换和均匀采样，在此稍微分析一下。常数Q变换公式通常如下：

* ：频率点的索引，通常从0开始。在CQT中，kk 用于表示在对数尺度上的位置。
* ：第 k个频率点在常数Q尺度上的值。这是变换后的频率，用于后续的信号分析。
* ：采样率，即信号每秒的采样次数。这个值决定了信号的最高频率（根据奈奎斯特定理，最高频率为采样率的一半）。
* ：参考频率，通常取值为0或采样率的一半。这个值用于确定变换的起始点。
* ：常数，决定了频率间隔的对数尺度。当 B 取值为1时，变换在Mel频率尺度上非常均匀。Mel频率尺度是基于人耳听觉特性的一种对数频率尺度，它在低频区域提供较宽的间隔，在高频区域提供较窄的间隔，以更好地模拟人耳对不同频率的敏感度。

可以观察到这个指数函数在 *k* 的值较小的时候增长得较慢，这意味着在低频区域，频率间隔较小，从而提供了更细的分辨率。所以它的设计想法与PLP是一致的，只是采用的公式不同，这个公式更加地简单些。

至于均匀采样则是确保了在频域分析中，每个频带的宽度是相同的。这有助于在不同频率区域中保持分析的一致性，使得特征提取不受频率分布的影响。

# 三、特征提取工具

librosa、Matlab和Kaldi都是在语音处理和语音识别领域中广泛使用的工具。它们都提供了多种音频特征提取方法，其中包括MFCC、FBank等常用的特征提取方法。

librosa是一个基于Python的开源库，主要用于音频分析和处理。它提供了多种特征提取方法，例如MFCC、FBank、Mel-Spectrogram等，并且具有良好的可视化效果，方便用户进行音频分析和处理。

Matlab是一种常用的数学计算软件，提供了许多用于语音处理和信号处理的工具箱和函数，例如Signal Processing Toolbox、Audio Toolbox和Wavelet Toolbox等。用户可以使用Matlab实现MFCC、FBank、LPC和其他语音特征提取。

Kaldi是一个基于C++的开源语音识别工具包，包含了语音识别的所有组件，包括前端特征提取、声学模型训练、解码器等。Kaldi支持多种特征提取方法，包括MFCC、FBank等，并且提供了丰富的工具和函数库，方便用户进行语音识别的开发和研究。

这些工具和库都可以帮助用户实现音频处理和语音识别的任务，并且具有不同的特点和优势，用户可以根据自己的需求选择合适的工具和方法。



图17 librosa和Matlab获取语音特征