Lecture 03: 语音信号处理基础

预处理与特征提取

王帅

南京大学智能科学与技术学院

2025 年 9 月 8 日

课程大纲

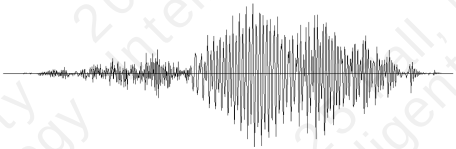
1. 引言与基础
2. 语音信号预处理 (Pre-processing)
3. 核心：特征提取 (Feature Extraction)
4. 总结 (Summary)

物理世界的语音信号

核心目标：从物理世界到数字世界

物理世界的语音信号

- 连续的压力波
- 模拟信号 (Analog)
- 信息丰富但高度冗余

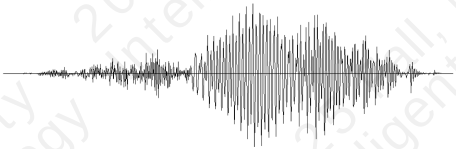


计算机世界的数字表示

核心目标：从物理世界到数字世界

物理世界的语音信号

- 连续的压力波
- 模拟信号 (Analog)
- 信息丰富但高度冗余



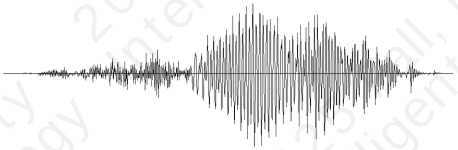
计算机世界的数字表示

- 离散的数值序列
- 数字信号 (Digital)
- 精炼、可计算的 特征

核心目标：从物理世界到数字世界

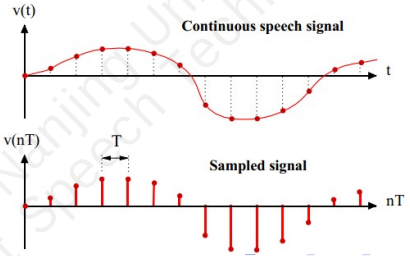
物理世界的语音信号

- 连续的压力波
- 模拟信号 (Analog)
- 信息丰富但高度冗余



计算机世界的数字表示

- 离散的数值序列
- 数字信号 (Digital)
- 精炼、可计算的 特征

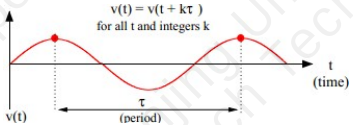


信号的类型

周期性

确定信号：信号根据已有的公式而产生

周期信号：根据周期 τ 进行重复



非周期信号：任何没有固定周期的信号

随机信号：在 t 时刻的信号是一个随机变量的函数：不可预见

人类产生的声波信号

发声原理

声源系统

呼吸系统提供的气流
喉部声带的振动

滤波器系统

声道（咽腔、口腔、鼻腔）
改变声音频谱特性

最终语音

我们听到的声音
包含共振峰、音高、音色

声源系统

呼吸系统
肺部作为气泵，通过膈肌和肋间肌的运动提供动力气流

喉部结构
声带（Vocal Folds）是发声的关键器官，两片富有弹性的肌肉膜

声带振动
气流通过声带时，声带周期性振动

滤波器系统

声道结构
可变的谐振腔，从喉部上方延伸至嘴唇和鼻孔，包括咽腔、口腔和鼻腔

共振峰（Formants）
声道内部空气柱发生共振，在特定频率上产生能量峰值

声道形状变化
不同形状导致不同共振峰频率，是区分不同音素的关键机制

人类产生的声波信号

发声原理

浊音(Voiced Sound)

声带处于适当张力并相互靠近，气流通过时引起声带周期性振动周期性振动产生基频 (F_0)

示例：元音 (a, e, i, o, u)、辅音 (m, n, b, d, g, z, v)

轻音(Unvoiced Sound)

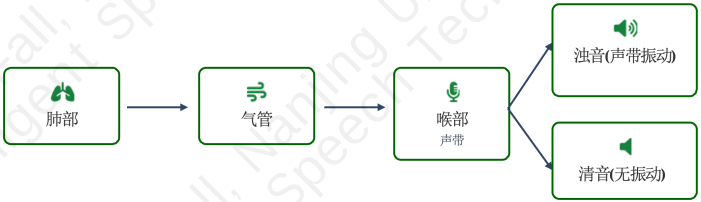
声带张开，气流通过时不会引起声带振动，直接通过喉部形成湍流无明显的基频，声学特性更接近噪声

示例：辅音 (p, t, k, f, s, h)

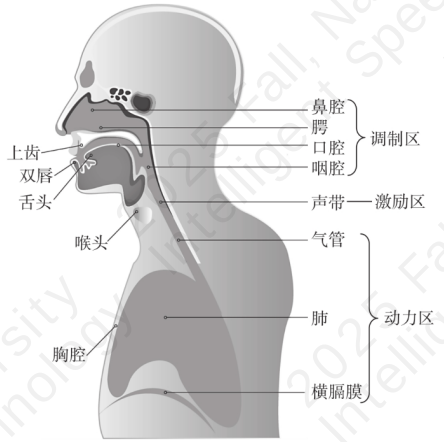
基频(Fundamental Frequency, F_0)

基频决定了语音的音高，声带振动越快，音高越高

↑成人男性基频范围：约80-120 Hz
↑成人女性基频范围：约160-240 Hz



人类的发音是一个复杂而精妙的过程，主要涉及动力区、激励区和调制区三个部分的器官协同作用。



调制区由舌头、上齿、双唇、咽腔等组成。

- 舌头的前后位置、高低位置以及卷曲程度的不同，可以产生不同的元音和辅音。
- 上齿和双唇可以通过开合、紧闭、突出等动作，与气流相互作用，产生不同的爆破音、摩擦音和鼻音等辅音。

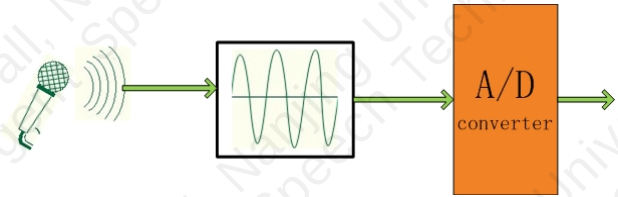
激励区包括喉头、气管和声带。

- 喉头又称喉结，由软骨、肌肉和韧带组成，内部有声带
- 气流从肺部经气管上升至喉头，声带根据发音需要振动或不振动

动力区主要由胸腔、肺和横膈膜组成。

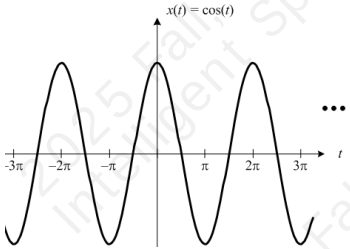
- 肺是呼吸主要器官，发音时肺部呼出气流提供动力
- 横膈膜位于胸腔和腹腔之间，在发音过程中，通过收缩和放松来调节气流的强弱和稳定性。

1. 清音段：能量低，过零率高，波形特点有点像随机的噪声。这部分信号常与语音的辅音段对应。
2. 浊音段：能量高，过零率低，波形具有周期性特点。所谓的短时平稳性质就是处于这个语音浊音（元音）段中。
3. 过渡段：一般是指从辅音段向元音段信号变化之间的部分。信号变化快，是语音信号处理中最复杂、困难的部分。

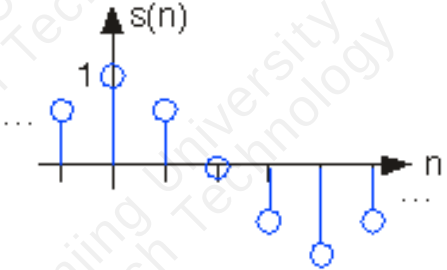


当我们对着一个麦克风说话的时候，声压的变化被转化成电压层面的成比例的变化。一台装配有合适硬件的计算机通过一个被称为模数转换 (ADC) 的过程，可以将模拟电压信号变化转化成数字声音的波形信号。

连续-时间/幅值信号:



离散-时间/幅值信号:



第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

$$f_s \geq 2 \cdot f_{max}$$

第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

$$f_s \geq 2 \cdot f_{max}$$

应用实例

- 人声有效频率范围：主要能量集中在 4 kHz 以下，但泛音可达更高。

第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

$$f_s \geq 2 \cdot f_{max}$$

应用实例

- 人声有效频率范围：主要能量集中在 4 kHz 以下，但泛音可达更高。
- 电话语音（窄带）：$f_{max} \approx 4$ kHz，因此行业标准采样率为 8 kHz。

第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

$$f_s \geq 2 \cdot f_{max}$$

应用实例

- 人声有效频率范围：主要能量集中在 4 kHz 以下，但泛音可达更高。
- 电话语音 (窄带)：$f_{max} \approx 4$ kHz，因此行业标准采样率为 8 kHz。
- 通用语音识别 (宽带)：为保留更丰富的泛音细节，通常 $f_{max} \approx 8$ kHz，标准采样率为 16 kHz。

第一步：数字化之采样 (Sampling)

时间维度的离散化

奈奎斯特-香农采样定理 (Nyquist-Shannon Sampling Theorem)

为了无失真地从样本中恢复原始模拟信号，采样频率 f_s 必须至少是原始信号最高频率 f_{max} 的两倍。

$$f_s \geq 2 \cdot f_{max}$$

应用实例

- 人声有效频率范围：主要能量集中在 4 kHz 以下，但泛音可达更高。
- 电话语音 (窄带)：$f_{max} \approx 4$ kHz，因此行业标准采样率为 8 kHz。
- 通用语音识别 (宽带)：为保留更丰富的泛音细节，通常 $f_{max} \approx 8$ kHz，标准采样率为 16 kHz。
- 关键问题：如果采样率不足 ($f_s < 2f_{max}$) 会发生什么？→ 频谱混叠 (Aliasing)。

第一步：数字化之量化 (Quantization)

幅度维度的离散化

- 采样后，每个样本点的幅度值是连续的。

第一步：数字化之量化 (Quantization)

幅度维度的离散化

- 采样后，每个样本点的幅度值是连续的。
- 量化 (Quantization) 是用有限个离散电平来近似表示连续的幅度值。

第一步：数字化之量化 (Quantization)

幅度维度的离散化

- 采样后，每个样本点的幅度值是连续的。
- 量化 (Quantization) 是用有限个离散电平来近似表示连续的幅度值。
- 使用 B 个比特 (bit) 表示一个样本，则共有 2^B 个量化级别。

第一步：数字化之量化 (Quantization)

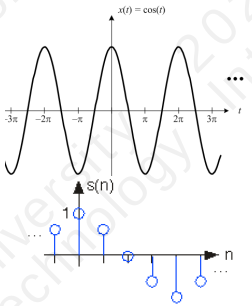
幅度维度的离散化

- 采样后，每个样本点的幅度值是连续的。
- 量化 (Quantization) 是用有限个离散电平来近似表示连续的幅度值。
- 使用 B 个比特 (bit) 表示一个样本，则共有 2^B 个量化级别。

第一步：数字化之量化 (Quantization)

幅度维度的离散化

- 采样后，每个样本点的幅度值是连续的。
- 量化 (Quantization) 是用有限个离散电平来近似表示连续的幅度值。
- 使用 B 个比特 (bit) 表示一个样本，则共有 2^B 个量化级别。



行业标准：16-bit PCM

16-bit 脉冲编码调制 **(PCM)** 是语音处理中最常用的标准。

- $B = 16$，提供 $2^{16} = 65,536$ 个量化级别。

量化误差 (Quantization Error)

$$e[n] = x_q[n] - x[n]$$

原始采样值 $x[n]$ 与量化后的值 $x_q[n]$ 的差。

核心假设：语音信号的短时平稳性 (Quasi-stationary)

语音信号在宏观上是时变的（非平稳），但在一个足够短的时间窗口内（如 10 ms to 30 ms），其发声器官状态和信号统计特性可以认为是基本不变的（平稳）。

预处理：为特征提取做准备

核心假设：语音信号的短时平稳性 (Quasi-stationary)

语音信号在宏观上是时变的（非平稳），但在一个足够短的时间窗口内（如 10 ms to 30 ms），其发声器官状态和信号统计特性可以认为是基本不变的（平稳）。

为什么？

我们的发声器官（声带、舌头、嘴唇）的物理状态在短时间内不会发生剧烈变化。

怎么办？

将信号切分成很多短小的“帧”(Frame)，对每一帧进行独立分析。这就是 分帧 的思想。

预处理：为特征提取做准备

核心假设：语音信号的短时平稳性 (Quasi-stationary)

语音信号在宏观上是时变的（非平稳），但在一个足够短的时间窗口内（如 10 ms to 30 ms），其发声器官状态和信号统计特性可以认为是基本不变的（平稳）。

为什么？

我们的发声器官（声带、舌头、嘴唇）的物理状态在短时间内不会发生剧烈变化。

怎么办？

将信号切分成很多短小的“帧”(Frame)，对每一帧进行独立分析。这就是 分帧 的思想。

经典预处理流程

去直流/预加重 → 分帧 → 加窗

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。

预处理 (1)

预加重 (Pre-emphasis)

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。
- 突出高频共振峰，这对区分不同的辅音（如 /s/, /f/）至关重要。

预处理 (1)

预加重 (Pre-emphasis)

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。
- 突出高频共振峰，这对区分不同的辅音（如 /s/, /f/）至关重要。
- 某种程度上，可以平滑频谱，并改善后续傅里叶变换的数值稳定性。

预处理 (1)

预加重 (Pre-emphasis)

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。
- 突出高频共振峰，这对区分不同的辅音（如 /s/, /f/）至关重要。
- 某种程度上，可以平滑频谱，并改善后续傅里叶变换的数值稳定性。

预处理 (1)

预加重 (Pre-emphasis)

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。
- 突出高频共振峰，这对区分不同的辅音（如 /s/, /f/）至关重要。
- 某种程度上，可以平滑频谱，并改善后续傅里叶变换的数值稳定性。

实现方法：一阶高通滤波器

将信号通过一个简单的一阶有限冲激响应 (FIR) 滤波器：

$$y[n] = x[n] - \alpha \cdot x[n-1]$$

预处理 (1)

预加重 (Pre-emphasis)

目的

- 提升高频分量的能量，补偿语音信号高频滚降的特性。
- 突出高频共振峰，这对区分不同的辅音（如 /s/, /f/）至关重要。
- 某种程度上，可以平滑频谱，并改善后续傅里叶变换的数值稳定性。

实现方法：一阶高通滤波器

将信号通过一个简单的一阶有限冲激响应 (FIR) 滤波器：

$$y[n] = x[n] - \alpha \cdot x[n-1]$$

- $y[n]$ 是预加重后的信号。
- $x[n]$ 是原始信号。
- α 是预加重系数，通常取值在 0.95 到 0.97 之间。

预处理 (2) & (3)

分帧与加窗

分帧 (Framing)

- 帧长 **(Frame Size):** 通常为 20 ms to 30 ms。(16 kHz 采样率下，25 ms = 400 个采样点)
- 帧移 **(Frame Shift):** 通常为 10 ms。(16 kHz 采样率下，10 ms = 160 个采样点)
- 帧与帧之间有 重叠，确保信息的平滑过渡。

加窗 (Windowing)

- 问题**:** 直接截断一帧信号（等于乘以矩形窗）会在频域引入严重的 频谱泄漏 (Spectral Leakage)。
- 方案**:** 乘以一个两端平滑过渡到零的窗函数，如汉明窗。

预处理 (2) & (3)

分帧与加窗

分帧 (Framing)

- 帧长 **(Frame Size):** 通常为 20 ms to 30 ms。(16 kHz 采样率下，25 ms = 400 个采样点)
- 帧移 **(Frame Shift):** 通常为 10 ms。(16 kHz 采样率下，10 ms = 160 个采样点)
- 帧与帧之间有 重叠，确保信息的平滑过渡。

加窗 (Windowing)

- 问题：直接截断一帧信号（等于乘以矩形窗）会在频域引入严重的 频谱泄漏 (Spectral Leakage)。
- 方案：乘以一个两端平滑过渡到零的窗函数，如汉明窗。

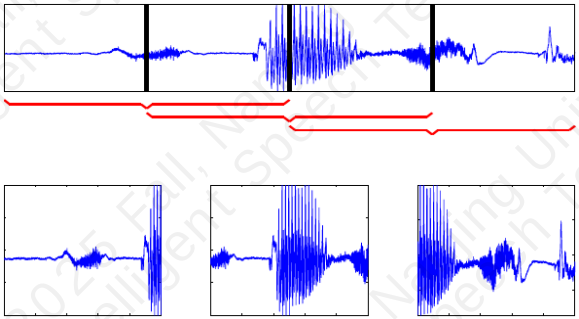
类比

直接分帧就像用剪刀“咔嚓”一下剪断绳子，断口会炸开；加窗则像是在剪断前，先用胶水把两端捻细，断口就会很平滑。

预处理 (2) & (3)

分帧

完整语音波形是一个很长非平稳长时采样序列. 在语音信号处理中，将长时序列分成准平稳的不同的块/帧是很有用的



汉明窗 (Hamming Window) - 语音领域最常用

对于一个长度为 N 的帧，窗函数 w[n] 定义为：

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

加窗后的信号 $x_w[n] = x[n] \cdot w[n]$。

汉明窗 (Hamming Window) - 语音领域最常用

对于一个长度为 N 的帧，窗函数 w[n] 定义为：

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

加窗后的信号 $x_w[n] = x[n] \cdot w[n]$。

矩形窗 (Rectangular)

- $w[n] = 1$
- 优点：主瓣窄 (频率分辨率高)
- 缺点：旁瓣高 (频谱泄漏严重)

汉明窗/汉宁窗

- 两端趋于 0
- 优点：旁瓣低 (有效抑制频谱泄漏)
- 缺点：主瓣略宽 (频率分辨率稍低)

汉明窗 (Hamming Window) - 语音领域最常用

对于一个长度为 N 的帧，窗函数 w[n] 定义为：

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

加窗后的信号 $x_w[n] = x[n] \cdot w[n]$。

矩形窗 (Rectangular)

- $w[n] = 1$
- 优点：主瓣窄 (频率分辨率高)
- 缺点：旁瓣高 (频谱泄漏严重)

汉明窗/汉宁窗

- 两端趋于 0
- 优点：旁瓣低 (有效抑制频谱泄漏)
- 缺点：主瓣略宽 (频率分辨率稍低)

多数情况抑制频谱泄漏比追求极致的频率分辨率重要，因此汉明窗是标准选择。

为什么要进行时频分析？

从时域波形到频域特征

时域分析的局限性

时域波形缺乏直观的结构性。同一个音素（如 /a/），不同人说出来，甚至同一个人在不同时间说，其波形都可能千差万别。

思考：我们如何才能“看透”波形，找到其背后稳定不变的本质？

为什么要进行时频分析？

从时域波形到频域特征

时域分析的局限性

时域波形缺乏直观的结构性。同一个音素（如 /a/），不同人说出来，甚至同一个人在不同时间说，其波形都可能千差万别。

思考：我们如何才能“看透”波形，找到其背后稳定不变的本质？

分析频率成分

语音的本质特征隐藏在频率中。例如，元音的身份由几个关键的共振峰 **(Formants)** 的频率位置决定。这些频率特征比时域波形稳定得多。

目标：我们需要一个工具，能揭示信号在不同时间点的频率构成。

基本原理：语音的短时平稳性 (Short-Time Stationarity)

语音信号整体上是非平稳的，其统计特性随时间快速变化。但在一个足够短的时间窗口内（如 20-30ms），声道形状可视为近似不变，信号表现出准平稳特性。

核心工具：短时傅里叶变换 (STFT)

捕捉时变的频谱特性

基本原理：语音的短时平稳性 (Short-Time Stationarity)

语音信号整体上是非平稳的，其统计特性随时间快速变化。但在一个足够短的时间窗口内（如 20-30ms），声道形状可视为近似不变，信号表现出准平稳特性。

STFT：滑动窗口分析法

STFT 的过程就像用一个“探照灯”（窗函数）在时间轴上滑动：

1. 分帧 (Framing): 将信号切分成一系列有重叠的短时帧。

声谱图是语音信号的“指纹”，可视化了能量在时间和频率上的分布。

核心工具：短时傅里叶变换 (STFT)

捕捉时变的频谱特性

基本原理：语音的短时平稳性 (Short-Time Stationarity)

语音信号整体上是非平稳的，其统计特性随时间快速变化。但在一个足够短的时间窗口内（如 20-30ms），声道形状可视为近似不变，信号表现出准平稳特性。

STFT：滑动窗口分析法

STFT 的过程就像用一个“探照灯”（窗函数）在时间轴上滑动：

1. 分帧 (Framing): 将信号切分成一系列有重叠的短时帧。
2. 加窗 (Windowing): 对每一帧乘以一个窗函数，以平滑帧的边缘。

声谱图是语音信号的“指纹”，可视化了能量在时间和频率上的分布。

核心工具：短时傅里叶变换 (STFT)

捕捉时变的频谱特性

基本原理：语音的短时平稳性 (Short-Time Stationarity)

语音信号整体上是非平稳的，其统计特性随时间快速变化。但在一个足够短的时间窗口内（如 20-30ms），声道形状可视为近似不变，信号表现出准平稳特性。

STFT：滑动窗口分析法

STFT 的过程就像用一个“探照灯”（窗函数）在时间轴上滑动：

1. 分帧 (Framing): 将信号切分成一系列有重叠的短时帧。
2. 加窗 (Windowing): 对每一帧乘以一个窗函数，以平滑帧的边缘。
3. 变换 (Transform): 对加窗后的每一帧进行傅里叶变换 (FFT)，得到该时间点的频谱。

声谱图是语音信号的“指纹”，可视化了能量在时间和频率上的分布。

核心工具：短时傅里叶变换 (STFT)

捕捉时变的频谱特性

基本原理：语音的短时平稳性 (Short-Time Stationarity)

语音信号整体上是非平稳的，其统计特性随时间快速变化。但在一个足够短的时间窗口内（如 20-30ms），声道形状可视为近似不变，信号表现出准平稳特性。

STFT：滑动窗口分析法

STFT 的过程就像用一个“探照灯”（窗函数）在时间轴上滑动：

1. 分帧 (Framing): 将信号切分成一系列有重叠的短时帧。
2. 加窗 (Windowing): 对每一帧乘以一个窗函数，以平滑帧的边缘。
3. 变换 (Transform): 对加窗后的每一帧进行傅里叶变换 (FFT)，得到该时间点的频谱。
4. 组合 (Combine): 将所有帧的频谱按时间顺序排列，形成声谱图 (Spectrogram)。

声谱图是语音信号的“指纹”，可视化了能量在时间和频率上的分布。

STFT 公式

在时间点 m 和频率点 k 的 STFT 值 $X(m,k)$ 定义为：

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mH] \cdot w[n] \cdot e^{-j\frac{2\pi kn}{N_{FFT}}}$$

- $x[\cdot]$: 原始信号
- $w[n]$: 长度为 N 的窗函数
- H: 帧移 (Hop Length)
- N_{FFT}: FFT 计算点数 (通常 $N_{FFT} \geq N$)

- 语音是非平稳信号，但在短时间（约 20 ms–30 ms）内可近似平稳：声道形状在该尺度内变化缓慢。

- 语音是非平稳信号，但在短时间（约 20 ms–30 ms）内可近似平稳：声道形状在该尺度内变化缓慢。
- 元音等共振峰（Formants）对应声道共振，短时幅度谱能稳定刻画其位置；辅音/爆破体现在高频能量与瞬态结构。

Why STFT ?

语音的短时平稳性与听觉相关性

- 语音是非平稳信号，但在短时间（约 20 ms–30 ms）内可近似平稳：声道形状在该尺度内变化缓慢。
- 元音等共振峰（Formants）对应声道共振，短时幅度谱能稳定刻画其位置；辅音/爆破体现在高频能量与瞬态结构。
- 发声源-声道模型：声源（脉冲列/噪声）经声道滤波，短时谱近似为声道响应的幅度包络 → 借助 STFT 可提取 MFCC、Mel 频谱等特征。

Why STFT ?

语音的短时平稳性与听觉相关性

- 语音是非平稳信号，但在短时间（约 20 ms–30 ms）内可近似平稳：声道形状在该尺度内变化缓慢。
- 元音等共振峰（Formants）对应声道共振，短时幅度谱能稳定刻画其位置；辅音/爆破体现在高频能量与瞬态结构。
- 发声源-声道模型：声源（脉冲列/噪声）经声道滤波，短时谱近似为声道响应的幅度包络 → 借助 STFT 可提取 MFCC、Mel 频谱等特征。
- 连续时间/全局频域会把不同语音单元的时变特性“混在一起”，难以对齐音素级事件；STFT 提供时频局部化表示，便于 ASR、TTS、增强与分离。

Why STFT ?

语音的短时平稳性与听觉相关性

- 语音是非平稳信号，但在短时间（约 20 ms–30 ms）内可近似平稳：声道形状在该尺度内变化缓慢。
- 元音等共振峰（Formants）对应声道共振，短时幅度谱能稳定刻画其位置；辅音/爆破体现在高频能量与瞬态结构。
- 发声源-声道模型：声源（脉冲列/噪声）经声道滤波，短时谱近似为声道响应的幅度包络 → 借助 STFT 可提取 MFCC、Mel 频谱等特征。
- 连续时间/全局频域会把不同语音单元的时变特性“混在一起”，难以对齐音素级事件；STFT 提供时频局部化表示，便于 ASR、TTS、增强与分离。
- 计算与工程可行性：S-DFT/FFT 高效、成熟，逆变换（iSTFT）易于重建（配合重叠加法与窗修正）。

1. 连续时间傅里叶变换 (CTFT)

- 对象: 连续、非周期信号 $x(t)$。
- 特点: 得到连续、非周期的频谱 $X(f)$。是纯粹的理论分析工具。

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt$$

理论溯源：傅里叶变换家族

从连续到离散，从理论到计算

1. 连续时间傅里叶变换 (CTFT)

- 对象: 连续、非周期信号 $x(t)$。
- 特点: 得到连续、非周期的频谱 $X(f)$。是纯粹的理论分析工具。

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt$$

2. 离散时间傅里叶变换 (DTFT)

- 对象: 离散、无限长信号 $x[n]$ (由 $x(t)$ 采样得到)。
- 特点: 得到连续、但以 2π 为周期的频谱 $X(e^{j\omega})$。理论上完美，但无法直接在计算机上精确计算。

3. 离散傅里叶变换 (DFT) 与快速傅里叶变换 (FFT)

- 对象: 离散、有限长信号 $x[n]$ (对 $x[n]$ 进行截断/分帧)。
- 特点: 得到离散、有限长的频谱 $X[k]$。

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{N}}$$

- **FFT**: 是一种高效计算 DFT 的算法，将复杂度从 $O(N^2)$ 降至 $O(N \log N)$，使得 STFT 等分析方法在工程上成为可能。

核心思想

MFCC (Mel-Frequency Cepstral Coefficients) 是一种模仿 人耳听觉机理 的特征。关键在于：人耳对频率的感知是 非线性的，对低频声音的变化比高频更敏感。

核心思想

MFCC (Mel-Frequency Cepstral Coefficients) 是一种模仿 人耳听觉机理 的特征。关键在于：人耳对频率的感知是 非线性的，对低频声音的变化比高频更敏感。

MFCC 提取全流程

1. 预处理: 预加重、分帧、加窗
2. **FFT**: 计算每一帧的功率谱

核心思想

MFCC (Mel-Frequency Cepstral Coefficients) 是一种模仿 人耳听觉机理 的特征。关键在于：人耳对频率的感知是 非线性的，对低频声音的变化比高频更敏感。

MFCC 提取全流程

1. 预处理: 预加重、分帧、加窗
2. **FFT**: 计算每一帧的功率谱
3. 梅尔滤波器组: 将功率谱通过一组特殊的三角滤波器

MFCC (梅尔频率倒谱系数)

核心思想

MFCC (Mel-Frequency Cepstral Coefficients) 是一种模仿 人耳听觉机理 的特征。关键在于：人耳对频率的感知是 非线性的，对低频声音的变化比高频更敏感。

MFCC 提取全流程

1. 预处理: 预加重、分帧、加窗
2. **FFT**: 计算每一帧的功率谱
3. 梅尔滤波器组: 将功率谱通过一组特殊的三角滤波器
4. 对数能量: 计算每个滤波器输出的对数能量

MFCC (梅尔频率倒谱系数)

核心思想

MFCC (Mel-Frequency Cepstral Coefficients) 是一种模仿 人耳听觉机理 的特征。关键在于：人耳对频率的感知是 非线性的，对低频声音的变化比高频更敏感。

MFCC 提取全流程

1. 预处理: 预加重、分帧、加窗
2. **FFT**: 计算每一帧的功率谱
3. 梅尔滤波器组: 将功率谱通过一组特殊的三角滤波器
4. 对数能量: 计算每个滤波器输出的对数能量
5. **DCT**: 进行离散余弦变换，解相关并降维

梅尔尺度 (Mel Scale)

一个基于人耳感知建立的非线性频率尺度。它与物理频率 f (Hz) 的转换关系近似为：

$$\text{mel} = 2595 \log_{10}\left(1 + \frac{\text{f}}{700}\right)$$

梅尔尺度 (Mel Scale)

一个基于人耳感知建立的非线性频率尺度。它与物理频率 f (Hz) 的转换关系近似为：

$$\text{mel} = 2595 \log_{10}\left(1 + \frac{\text{f}}{700}\right)$$

- 特点：低频区域，梅尔尺度与赫兹尺度近似线性；高频区域近似对数关系。

梅尔尺度 (Mel Scale)

一个基于人耳感知建立的非线性频率尺度。它与物理频率 f (Hz) 的转换关系近似为：

$$\text{mel} = 2595 \log_{10}\left(1 + \frac{\text{f}}{700}\right)$$

- 特点：低频区域，梅尔尺度与赫兹尺度近似线性；高频区域近似对数关系。
- 滤波器组：一组三角滤波器（通常 20-40 个），在梅尔尺度上等距分布。

梅尔尺度 (Mel Scale)

一个基于人耳感知建立的非线性频率尺度。它与物理频率 f (Hz) 的转换关系近似为：

$$\text{mel} = 2595 \log_{10}\left(1 + \frac{\text{f}}{700}\right)$$

- 特点：低频区域，梅尔尺度与赫兹尺度近似线性；高频区域近似对数关系。
- 滤波器组：一组三角滤波器（通常 20-40 个），在梅尔尺度上等距分布。
- 结果：这意味着在 低频区滤波器密集、窄小，在 高频区滤波器稀疏、宽大，完美模拟人耳特性。

步骤 4: 计算对数能量

将功率谱与每个梅尔滤波器相乘再求和，得到每个滤波器的输出能量 E_i，然后取对数。

$$S[i] = \ln\left(\sum_k P[k]M_i[k]\right)$$

- $M_i[k]$ 是第 i 个梅尔滤波器。

步骤 4: 计算对数能量

将功率谱与每个梅尔滤波器相乘再求和，得到每个滤波器的输出能量 E_i，然后取对数。

$$S[i] = \ln\left(\sum_k P[k]M_i[k]\right)$$

- $M_i[k]$ 是第 i 个梅尔滤波器。
- 为何取对数？1. 模仿人耳对声音强度的对数感知；2. 压缩数值动态范围。

步骤 4: 计算对数能量

将功率谱与每个梅尔滤波器相乘再求和，得到每个滤波器的输出能量 E_i，然后取对数。

$$S[i] = \ln\left(\sum_k P[k]M_i[k]\right)$$

- $M_i[k]$ 是第 i 个梅尔滤波器。
- 为何取对数？1. 模仿人耳对声音强度的对数感知；2. 压缩数值动态范围。

步骤 4: 计算对数能量

将功率谱与每个梅尔滤波器相乘再求和，得到每个滤波器的输出能量 E_i，然后取对数。

$$S[i] = \ln\left(\sum_k P[k]M_i[k]\right)$$

- $M_i[k]$ 是第 i 个梅尔滤波器。
- 为何取对数？1. 模仿人耳对声音强度的对数感知；2. 压缩数值动态范围。

步骤 5: 离散余弦变换 (DCT)

对数能量 S[i] 之间存在相关性。DCT 是一种优秀的正交变换，可以有效解相关，并将能量集中在少数低阶系数上。

$$c[n] = \sum_{i=1}^{K} S[i]\cos\left(\frac{\pi n(i-0.5)}{K}\right)$$

步骤 4: 计算对数能量

将功率谱与每个梅尔滤波器相乘再求和，得到每个滤波器的输出能量 E_i，然后取对数。

$$S[i] = \ln\left(\sum_k P[k]M_i[k]\right)$$

- $M_i[k]$ 是第 i 个梅尔滤波器。
- 为何取对数？1. 模仿人耳对声音强度的对数感知；2. 压缩数值动态范围。

步骤 5: 离散余弦变换 (DCT)

对数能量 S[i] 之间存在相关性。DCT 是一种优秀的正交变换，可以有效解相关，并将能量集中在少数低阶系数上。

$$c[n] = \sum_{i=1}^{K} S[i]\cos\left(\frac{\pi n(i-0.5)}{K}\right)$$

动态特征：Delta 与 Delta-Delta

为什么需要动态特征？

MFCC 是“静态”的，只描述了当前帧的频谱包络。但语音的精髓在于“变化”。例如，/b/ 和 /w/ 的稳态部分可能相似，但它们的 过渡（动态） 完全不同。

动态特征：Delta 与 Delta-Delta

为什么需要动态特征？

MFCC 是“静态”的，只描述了当前帧的频谱包络。但语音的精髓在于“变化”。例如，/b/ 和 /w/ 的稳态部分可能相似，但它们的 过渡（动态） 完全不同。

计算方法：差分 (Differentiation)

第 t 帧的 Delta 特征 d_t 可以通过一个回归窗口来近似计算其一阶导数：

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

- c_t 是第 t 帧的静态 MFCC 向量。N 是回归窗口大小 (通常为 2)。

动态特征：Delta 与 Delta-Delta

为什么需要动态特征？

MFCC 是“静态”的，只描述了当前帧的频谱包络。但语音的精髓在于“变化”。例如，/b/ 和 /w/ 的稳态部分可能相似，但它们的 过渡（动态） 完全不同。

计算方法：差分 (Differentiation)

第 t 帧的 Delta 特征 d_t 可以通过一个回归窗口来近似计算其一阶导数：

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

- c_t 是第 t 帧的静态 MFCC 向量。N 是回归窗口大小 (通常为 2)。
- 这本质上是在计算 MFCC 轨迹的 速度 **(velocity)**。

动态特征：Delta 与 Delta-Delta

为什么需要动态特征？

MFCC 是“静态”的，只描述了当前帧的频谱包络。但语音的精髓在于“变化”。例如，/b/ 和 /w/ 的稳态部分可能相似，但它们的 过渡（动态） 完全不同。

计算方法：差分 (Differentiation)

第 t 帧的 Delta 特征 d_t 可以通过一个回归窗口来近似计算其一阶导数：

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

- c_t 是第 t 帧的静态 MFCC 向量。N 是回归窗口大小 (通常为 2)。
- 这本质上是在计算 MFCC 轨迹的 速度 **(velocity)**。

动态特征：Delta 与 Delta-Delta

为什么需要动态特征？

MFCC 是“静态”的，只描述了当前帧的频谱包络。但语音的精髓在于“变化”。例如，/b/ 和 /w/ 的稳态部分可能相似，但它们的 过渡（动态） 完全不同。

计算方法：差分 (Differentiation)

第 t 帧的 Delta 特征 d_t 可以通过一个回归窗口来近似计算其一阶导数：

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

- c_t 是第 t 帧的静态 MFCC 向量。N 是回归窗口大小 (通常为 2)。
- 这本质上是在计算 MFCC 轨迹的 速度 **(velocity)**。

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) $\rightarrow$ 数字波形

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：

→ 加窗后的信号帧序列

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)

→ 加窗后的信号帧序列

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)

→ 加窗后的信号帧序列

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**):**

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**)**：
   - FFT → 功率谱

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**)**：
   - FFT → 功率谱
   - 梅尔滤波器组 → 对数能量

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**)**：
   - FFT → 功率谱
   - 梅尔滤波器组 → 对数能量
   - DCT → 静态 **MFCC** 特征

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**)**：
   - FFT → 功率谱
   - 梅尔滤波器组 → 对数能量
   - DCT → 静态 **MFCC** 特征
5. 动态特征计算：

总结：完整的信号处理流水线

1. 模拟信号 $\xrightarrow{\text{麦克风}}$
2. 数字化：采样 (16 kHz) & 量化 (16-bit) → 数字波形
3. 预处理：
   - 预加重 ($\alpha \approx 0.97$)
   - 分帧 (25 ms 帧长, 10 ms 帧移)
   - 加窗 (汉明窗)

   → 加窗后的信号帧序列
4. 特征提取 **(**以 **MFCC** 为例**)**：
   - FFT → 功率谱
   - 梅尔滤波器组 → 对数能量
   - DCT → 静态 **MFCC** 特征
5. 动态特征计算：
   - 差分 → **Delta, Delta-Delta** 特征

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

下一步是什么？

这些特征向量将作为输入，送入各种强大的机器学习模型中进行声学建模，例如：

- 传统模型：GMM，HMM (高斯混合模型，隐马尔可夫模型)

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

下一步是什么？

这些特征向量将作为输入，送入各种强大的机器学习模型中进行声学建模，例如：

- 传统模型：GMM，HMM (高斯混合模型，隐马尔可夫模型)
- 深度学习模型：

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

下一步是什么？

这些特征向量将作为输入，送入各种强大的机器学习模型中进行声学建模，例如：

- 传统模型：GMM，HMM (高斯混合模型，隐马尔可夫模型)
- 深度学习模型：
  - DNN, CNN, RNN/LSTM, Transformer...

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

下一步是什么？

这些特征向量将作为输入，送入各种强大的机器学习模型中进行声学建模，例如：

- 传统模型：GMM，HMM (高斯混合模型，隐马尔可夫模型)
- 深度学习模型：
  - DNN, CNN, RNN/LSTM, Transformer...

我们得到了什么？

我们成功地将一维、冗长、复杂的原始声波，转化为了一个低维、信息密集、相对鲁棒的特征向量序列。

$$\text{信号处理} \Longrightarrow \text{特征} \Longrightarrow \text{机器学习模型}$$

下一步是什么？

这些特征向量将作为输入，送入各种强大的机器学习模型中进行声学建模，例如：

- 传统模型：GMM，HMM (高斯混合模型，隐马尔可夫模型)
- 深度学习模型：
  - DNN, CNN, RNN/LSTM, Transformer...

这些模型将学习从特征到音素（或其他声学单元）的映射关系，从而实现各种任务。

- 深度学习把“特征提取器”与“任务头”合并到一个可微模型中：原始波形 → 多层时频卷积/注意力 → 任务。

思考：端到端系统的崛起

从手工特征到表征学习的范式迁移

- 深度学习把“特征提取器”与“任务头”合并到一个可微模型中：原始波形 → 多层时频卷积/注意力 → 任务。
- 自监督表征在低标注/跨任务迁移上展现压倒性优势。

思考：端到端系统的崛起

从手工特征到表征学习的范式迁移

- 深度学习把“特征提取器”与“任务头”合并到一个可微模型中：原始波形 → 多层时频卷积/注意力 → 任务。
- 自监督表征在低标注/跨任务迁移上展现压倒性优势。
- 但“特征工程”并未消失，而是上移为归纳偏置：窗口化、卷积核、频带分解、数据增强、目标函数、架构形态本身就是“可学习的特征框架”。

思考：端到端系统的崛起

从手工特征到表征学习的范式迁移

- 深度学习把“特征提取器”与“任务头”合并到一个可微模型中：原始波形 → 多层时频卷积/注意力 → 任务。
- 自监督表征在低标注/跨任务迁移上展现压倒性优势。
- 但“特征工程”并未消失，而是上移为归纳偏置：窗口化、卷积核、频带分解、数据增强、目标函数、架构形态本身就是“可学习的特征框架”。
- 经验律：数据越大、任务越复杂、部署资源越充分，端到端收益越明显；数据/算力受限时，合适的先验（如 Mel/STFT）可显著提升样本效率与鲁棒性。

- 样本效率：在小数据场景，显式 STFT/Mel 引入强先验（短时平稳、感知频带）能减少假设空间，训练更稳；Raw 端到端易过拟合，需要强正则与数据扩增。

Raw waveform vs. 频域先验

样本效率、鲁棒性、可解释性三角权衡

- 样本效率：在小数据场景，显式 STFT/Mel 引入强先验（短时平稳、感知频带）能减少假设空间，训练更稳；Raw 端到端易过拟合，需要强正则与数据扩增。
- 噪声鲁棒：频域滤波/子带建模易注入频率不变性与带噪可分性；Raw 端到端可学到更细粒度时域线索，但需更大数据来覆盖多种噪声/说话人/通道域。

Raw waveform vs. 频域先验

样本效率、鲁棒性、可解释性三角权衡

- 样本效率：在小数据场景，显式 STFT/Mel 引入强先验（短时平稳、感知频带）能减少假设空间，训练更稳；Raw 端到端易过拟合，需要强正则与数据扩增。
- 噪声鲁棒：频域滤波/子带建模易注入频率不变性与带噪可分性；Raw 端到端可学到更细粒度时域线索，但需更大数据来覆盖多种噪声/说话人/通道域。
- 可解释性与可控性：显式谱图利于可视化、调参与诊断（泄漏、过平滑、带限）；Raw 输入强但“黑箱化”，调优靠大规模验证与可解释工具（CAM、频带敏感度分析）。

Raw waveform vs. 频域先验

样本效率、鲁棒性、可解释性三角权衡

- 样本效率：在小数据场景，显式 STFT/Mel 引入强先验（短时平稳、感知频带）能减少假设空间，训练更稳；Raw 端到端易过拟合，需要强正则与数据扩增。
- 噪声鲁棒：频域滤波/子带建模易注入频率不变性与带噪可分性；Raw 端到端可学到更细粒度时域线索，但需更大数据来覆盖多种噪声/说话人/通道域。
- 可解释性与可控性：显式谱图利于可视化、调参与诊断（泄漏、过平滑、带限）；Raw 输入强但“黑箱化”，调优靠大规模验证与可解释工具（CAM、频带敏感度分析）。
- 推理代价：端到端可省显式特征计算（或以可学习前端替代），但整体模型更大；在边缘设备上，轻量显式特征 + 小模型仍具工程优势。

- 可学习滤波器组：用参数化卷积核初始化为 **Gammatone/Mel** 滤波器，再端到端微调，兼得先验与适应性。

- 可学习滤波器组：用参数化卷积核初始化为 **Gammatone/Mel** 滤波器，再端到端微调，兼得先验与适应性。
- 可学习 STFT：把 STFT 看作固定线性投影（余弦/正弦字典）；将其推广为可训练字典或多分辨率字典（多窗长、多采样 hop），提升时频自适应。

- 可学习滤波器组：用参数化卷积核初始化为 **Gammatone/Mel** 滤波器，再端到端微调，兼得先验与适应性。
- 可学习 STFT：把 STFT 看作固定线性投影（余弦/正弦字典）；将其推广为可训练字典或多分辨率字典（多窗长、多采样 hop），提升时频自适应。
- 复数域/相位感知：显式在复数 STFT 上建模幅度与相位，或学习相位敏感前端（在增强与分离中提升清晰度与可重建性）。

可学习前端：连接传统先验与端到端

从固定 STFT/Mel 到可训练滤波器组

- 可学习滤波器组：用参数化卷积核初始化为 **Gammatone/Mel** 滤波器，再端到端微调，兼得先验与适应性。
- 可学习 STFT：把 STFT 看作固定线性投影（余弦/正弦字典）；将其推广为可训练字典或多分辨率字典（多窗长、多采样 hop），提升时频自适应。
- 复数域/相位感知：显式在复数 STFT 上建模幅度与相位，或学习相位敏感前端（在增强与分离中提升清晰度与可重建性）。
- 子带建模：把宽带任务拆成并行子带（变换域或可学习滤组），降低建模难度与计算，常见于高保真 TTS/音频生成。

- 表征的关键不止“输入长什么样”，更在于训练目标：对比学习、遮挡预测、跨模态对齐（语音-文本-说话人-情感）在很大程度决定“特征质量”。

- 表征的关键不止“输入长什么样”，更在于训练目标：对比学习、遮挡预测、跨模态对齐（语音-文本-说话人-情感）在很大程度决定“特征质量”。
- 数据配方即强归纳偏置：多域混合（远讲/近讲、房间/户外）、语言/口音多样性、设备多样性，是鲁棒表征的核心“特征工程”。

自监督与大模型时代：特征的“新角色”

从输入表示到训练目标与数据配方

- 表征的关键不止“输入长什么样”，更在于训练目标：对比学习、遮挡预测、跨模态对齐（语音-文本-说话人-情感）在很大程度决定“特征质量”。
- 数据配方即强归纳偏置：多域混合（远讲/近讲、房间/户外）、语言/口音多样性、设备多样性，是鲁棒表征的核心“特征工程”。
- 适配与蒸馏：用大模型作“教师”，小模型蒸馏到轻量前端；或把自监督特征冻住作为可迁移输入，任务头轻量化微调，快速适配下游场景。

自监督与大模型时代：特征的“新角色”

从输入表示到训练目标与数据配方

- 表征的关键不止“输入长什么样”，更在于训练目标：对比学习、遮挡预测、跨模态对齐（语音-文本-说话人-情感）在很大程度决定“特征质量”。
- 数据配方即强归纳偏置：多域混合（远讲/近讲、房间/户外）、语言/口音多样性、设备多样性，是鲁棒表征的核心“特征工程”。
- 适配与蒸馏：用大模型作“教师”，小模型蒸馏到轻量前端；或把自监督特征冻住作为可迁移输入，任务头轻量化微调，快速适配下游场景。
- 多任务与指令化：ASR/说话人/语种/情感联合训练或指令式建模，让同一表征服务多个任务——“通用音频 backbone”成为新特征工程。

洞见：为何短时思想仍然长期有效？

不确定性原理、发声机理与多尺度表示

- 不确定性权衡：时间-频率分辨率的物理极限并未因深度学习消失；模型可以学到多尺度，但窗口化/多尺度栈仍是稳定高效的实现方式。

洞见：为何短时思想仍然长期有效？

不确定性原理、发声机理与多尺度表示

- 不确定性权衡：时间-频率分辨率的物理极限并未因深度学习消失；模型可以学到多尺度，但窗口化/多尺度栈仍是稳定高效的实现方式。
- 发声机理先验：声门脉冲列 + 声道共振是局部平稳的，短时谱天然对共振峰/谐波结构友好；端到端常隐式学到“近似 STFT”的第一层卷积核。

洞见：为何短时思想仍然长期有效？

不确定性原理、发声机理与多尺度表示

- 不确定性权衡：时间-频率分辨率的物理极限并未因深度学习消失；模型可以学到多尺度，但窗口化/多尺度栈仍是稳定高效的实现方式。
- 发声机理先验：声门脉冲列 + 声道共振是局部平稳的，短时谱天然对共振峰/谐波结构友好；端到端常隐式学到“近似 STFT”的第一层卷积核。
- 任务对齐：音素/子词级标签与短时窗匹配良好，便于强监督/CTC 对齐；完全时域端到端在对齐上往往需要更强的注意力与正则。

洞见：为何短时思想仍然长期有效？

不确定性原理、发声机理与多尺度表示

- 不确定性权衡：时间-频率分辨率的物理极限并未因深度学习消失；模型可以学到多尺度，但窗口化/多尺度栈仍是稳定高效的实现方式。
- 发声机理先验：声门脉冲列 + 声道共振是局部平稳的，短时谱天然对共振峰/谐波结构友好；端到端常隐式学到“近似 STFT”的第一层卷积核。
- 任务对齐：音素/子词级标签与短时窗匹配良好，便于强监督/CTC 对齐；完全时域端到端在对齐上往往需要更强的注意力与正则。
- 可逆与可控：STFT/iSTFT 保证能量守恒与可逆重建，是生成/增强/编辑的工业硬基座；可学习前端可在其上叠加，而非完全替代。

1. 资源与数据量评估：

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。

工程建议：如何在项目中做取舍？

三步决策树

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。
2. 目标与可解释性需求：

工程建议：如何在项目中做取舍？

三步决策树

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。
2. 目标与可解释性需求：
   - 可控生成/音质可诊断（TTS/后期）：保留频域可视化与可逆流程（iSTFT），必要时复数域建模。

工程建议：如何在项目中做取舍？

三步决策树

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。
2. 目标与可解释性需求：
   - 可控生成/音质可诊断（TTS/后期）：保留频域可视化与可逆流程（iSTFT），必要时复数域建模。
   - 纯识别/检索：可直接用大表征（SSL）嵌入，冻结或小步微调下游头。

工程建议：如何在项目中做取舍？

三步决策树

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。
2. 目标与可解释性需求：
   - 可控生成/音质可诊断（TTS/后期）：保留频域可视化与可逆流程（iSTFT），必要时复数域建模。
   - 纯识别/检索：可直接用大表征（SSL）嵌入，冻结或小步微调下游头。
3. 泛化与鲁棒：

工程建议：如何在项目中做取舍？

三步决策树

1. 资源与数据量评估：
   - 小数据/边缘设备/强实时：优先 STFT/Mel + 轻量模型（Conformer-lite/TCN）；善用数据增强（混响、噪声、SpecAugment）。
   - 大数据/服务器端：端到端原始波形或可学习前端 + 自监督预训练，再微调。
2. 目标与可解释性需求：
   - 可控生成/音质可诊断（TTS/后期）：保留频域可视化与可逆流程（iSTFT），必要时复数域建模。
   - 纯识别/检索：可直接用大表征（SSL）嵌入，冻结或小步微调下游头。
3. 泛化与鲁棒：
   - 跨域部署，优先多域自监督 + 多任务正则；保持前端多尺度或子带并行以缓解分布偏移。

小结：特征设计没有消失，它进化了

从手工参数到可学习先验

- 端到端不是“不要特征”，而是把特征变成可训练的前端与目标函数。
- STFT/短时思想仍是稳定、高效、可逆的工程支点；可学习滤组与多尺度建模把它升级为数据驱动的版本。
- 真正的“特征工程 2.0”：数据配方、自监督任务、架构归纳偏置、鲁棒性增强与蒸馏/适配。

- MFCC 教程
- 傅里叶变换“掐死”教程（着重看图示）