数字媒体技术基础

Meng Yang

www.smartllv.com

SUN YAT-SEN University

机器智能与先进计算教育部重点实验室

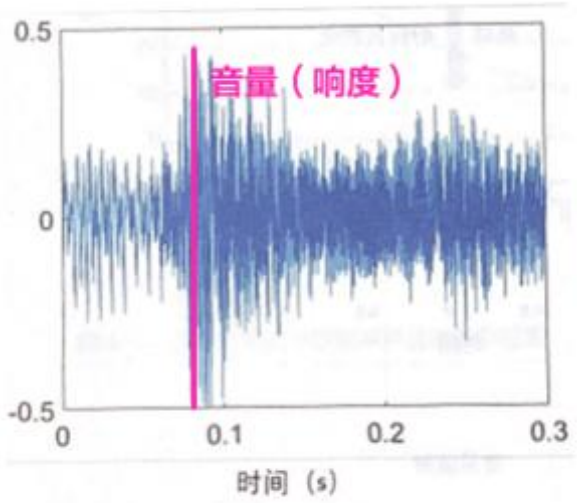
智能视觉语言学习研究组

目 录

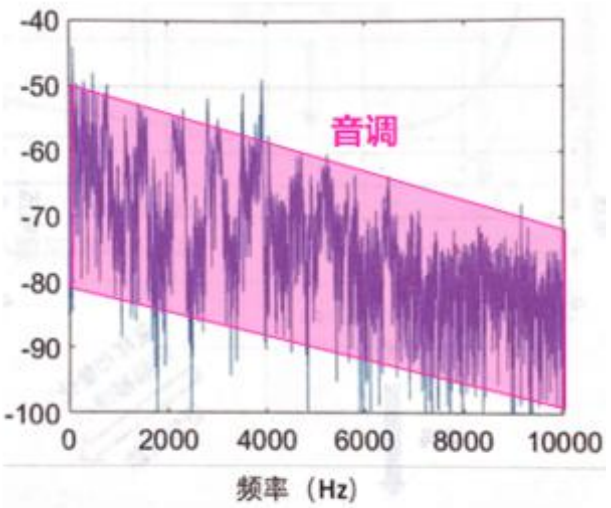
6 音频媒体信息表示

Course Outline

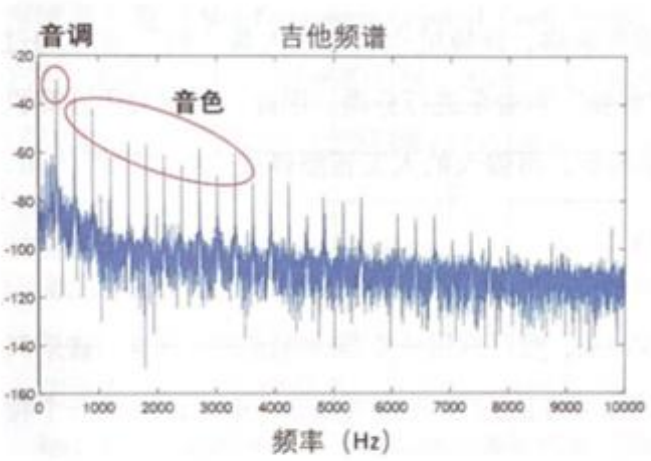
人类感官特征：音量、音调、音色

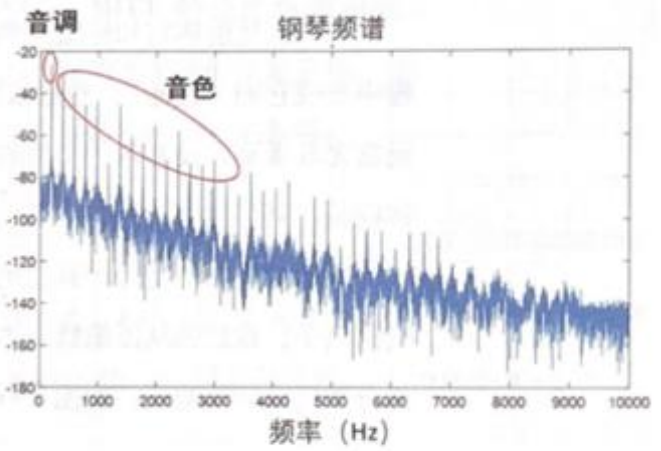


波形图



频率图





频率图

音色是指不同声音表现在波形方面总是有与众不同的特性，不同的物体振动都有不同的特点

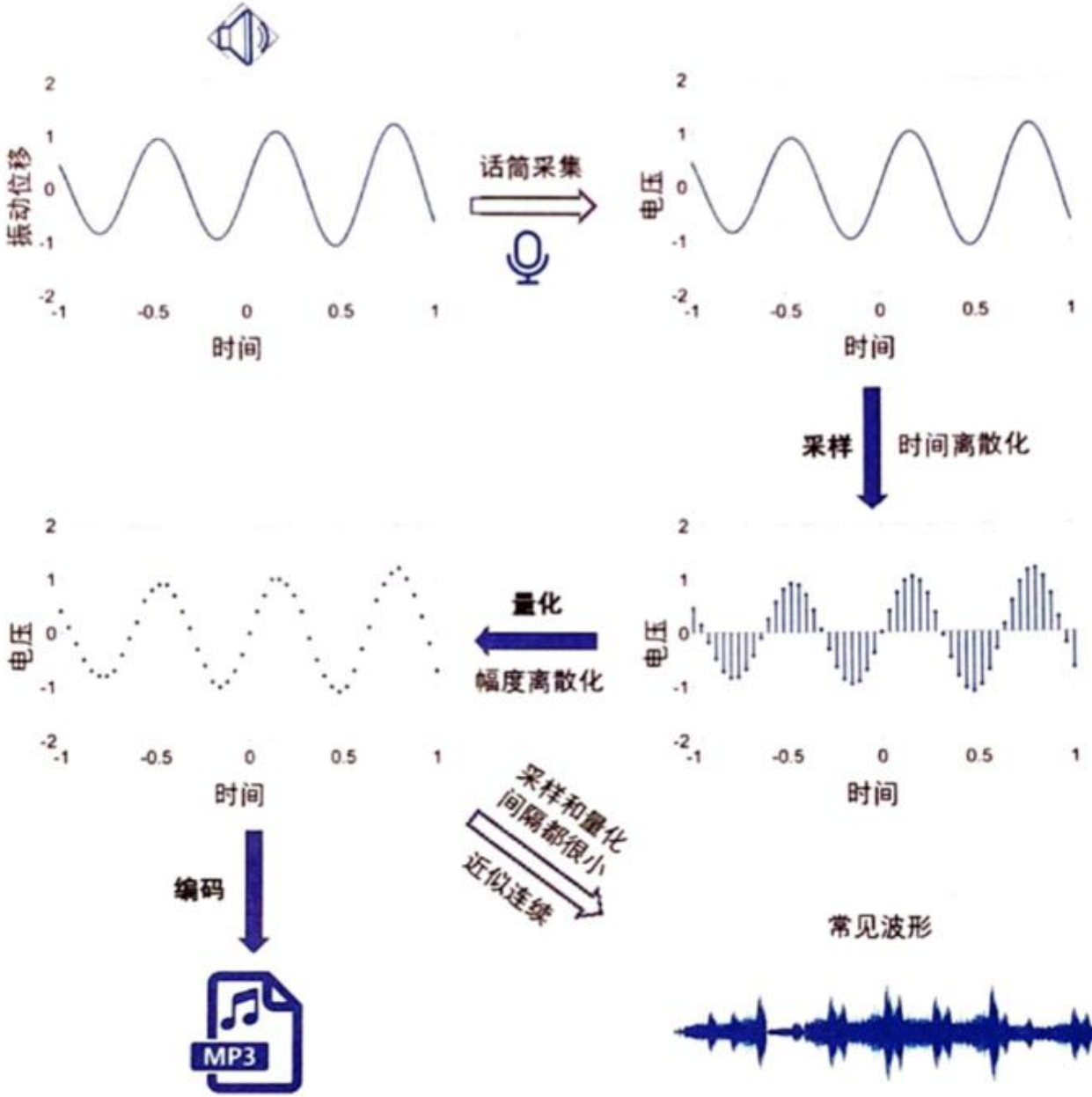
目 录

6.1 声音的数字化表示

6.1.1 声音的采样与量化

声转电

- 通过采样、
- 量化、
- 编码



6.1.1 声音的采样与量化

声音的数字化

- 声音信号

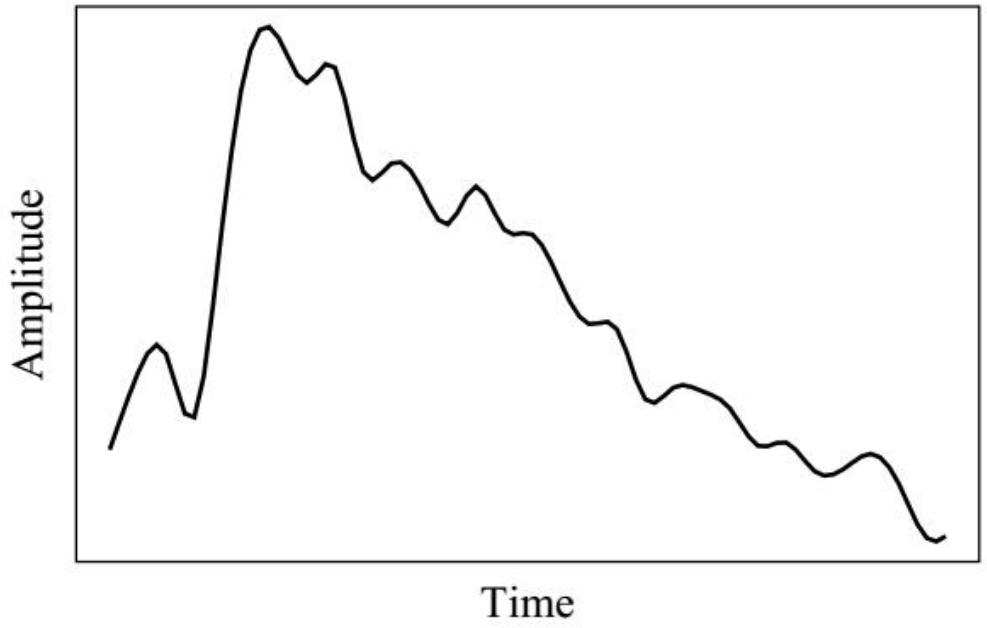


Fig. 6.1: An analog signal: continuous measurement of pressure wave.

6.1.1 声音的采样与量化

声音的数字化PCM（脉冲编码调制）方法

Pulse Code Modulation

- The basic techniques for creating digital signals from analog signals are **sampling** and **quantization**.
- Quantization consists of selecting breakpoints in magnitude, and then re-mapping any value within an interval to one of the representative output levels. $\longrightarrow$ Repeat of Fig. 6.2:

6.1.1 声音的采样与量化

采样与量化

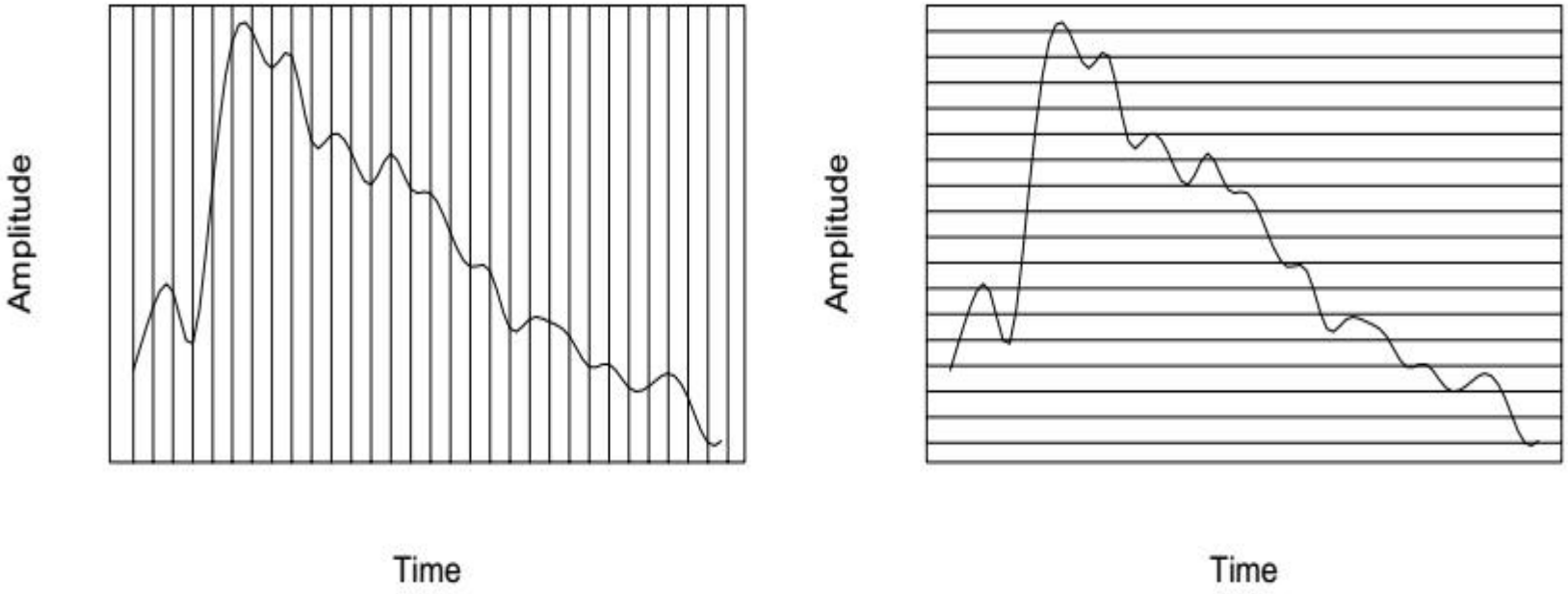


Fig. 6.2: Sampling and Quantization.

Weber's Law

- 问题：手机掉落的声音，为什么在嘈杂的地铁上不容易听到，而在安静的教室容易听到？

6.1.1 声音的采样与量化

声音的非线性量化变换函数

- **Non-uniform quantization**: set up more finely-spaced levels where humans hear with the most acuity.
  - Weber's Law stated formally says that equally perceived differences have values proportional to absolute levels:

$$\Delta\text{Response} \propto \Delta\text{Stimulus/Stimulus} \tag{6.5}$$

  - Inserting a constant of proportionality k, we have a differential equation that states:

$$dr = k\,(1/s)\,ds \tag{6.6}$$

    with response r and stimulus s.

6.1.1 声音的采样与量化

声音的非线性量化变换函数

– Integrating, we arrive at a solution

$$r = k \ln s + C \tag{6.7}$$

with constant of integration C.

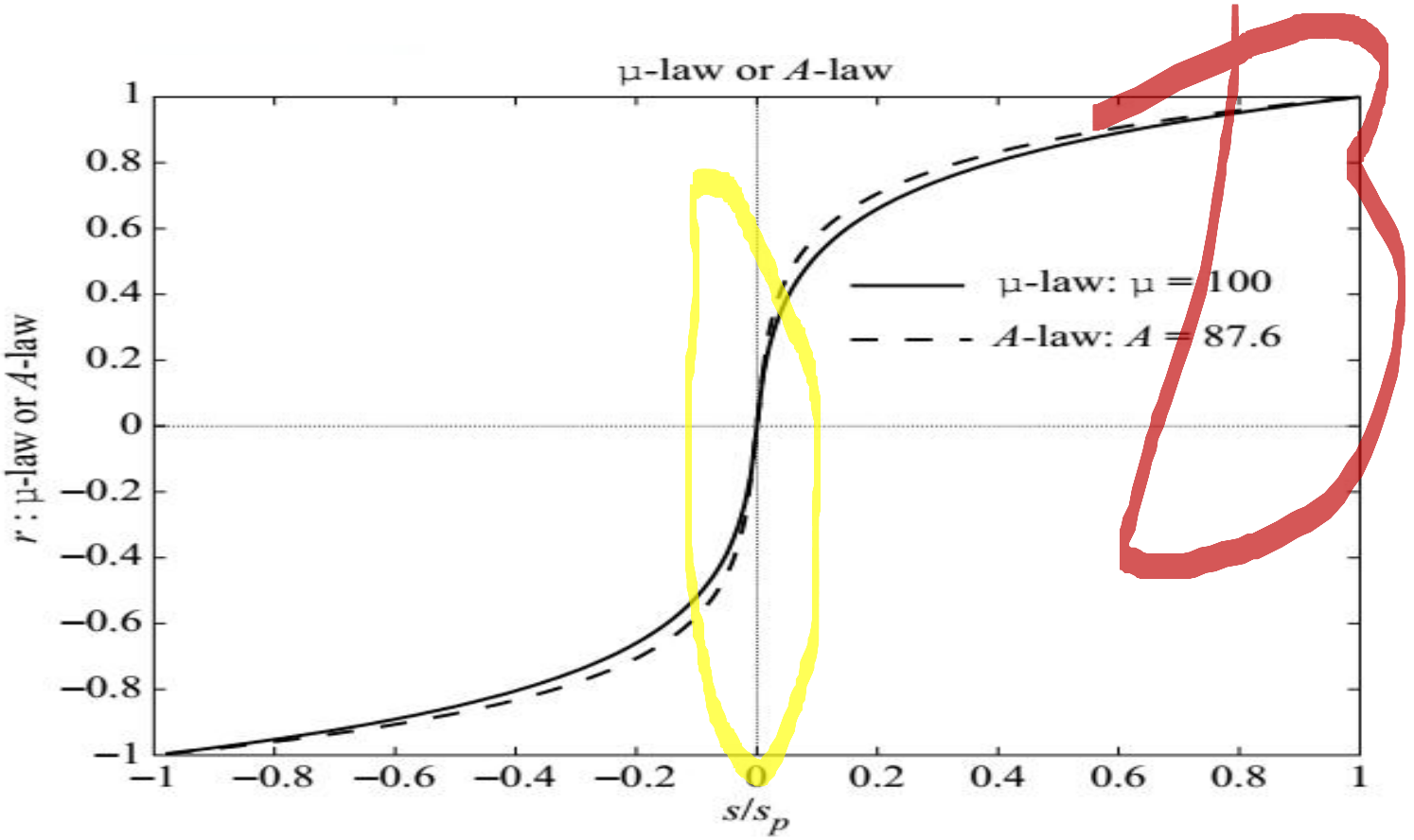
Stated differently, the solution is

$$r = k \ln(s/s_0) \tag{6.8}$$

s_0 = the lowest level of stimulus that causes a response ($r = 0$ when $s = s_0$).

6.1.1 声音的采样与量化

声音的非线性量化变换函数



ig. 6.6: Nonlinear transform for audio signals

6.1.1 声音的采样与量化

PCM过程

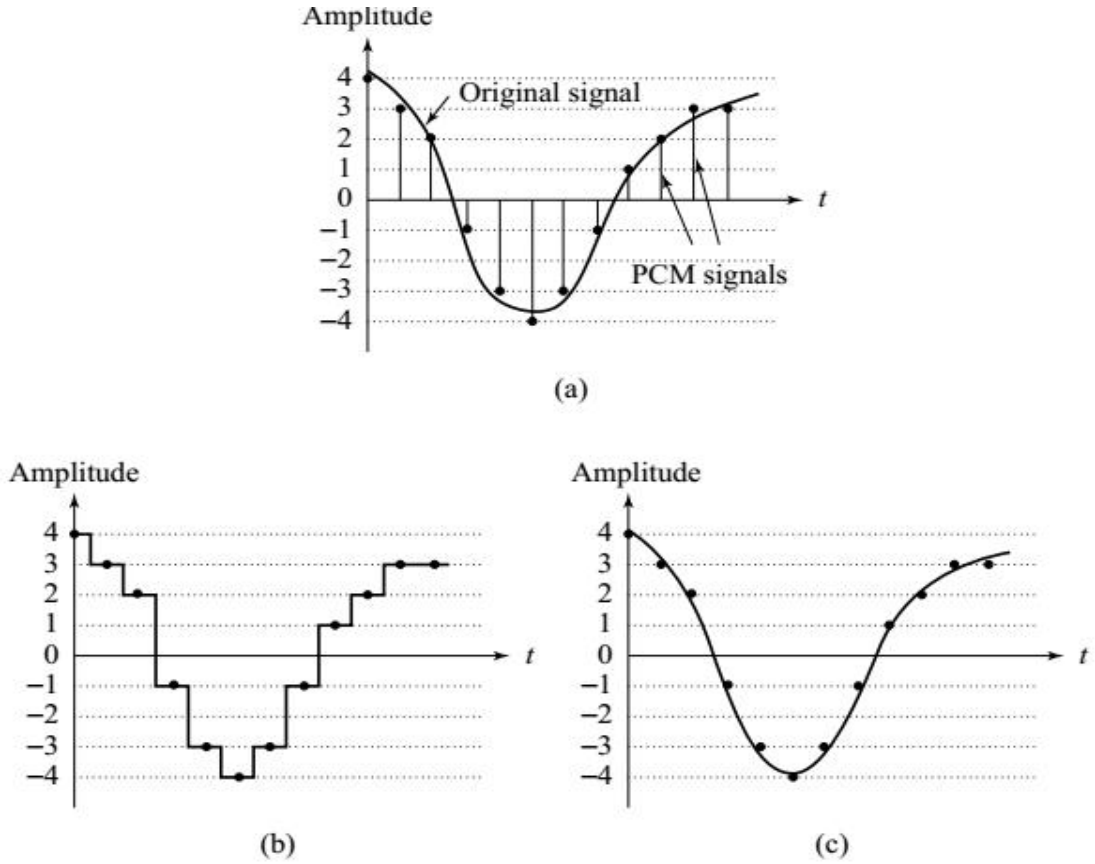


Fig. 6.13: Pulse Code Modulation (PCM). (a) Original analog signal and its corresponding PCM signals. (b) Decoded staircase signal. (c) Reconstructed signal after low-pass filtering.

6.1.1 声音的采样与量化

PCM编码和解码过程

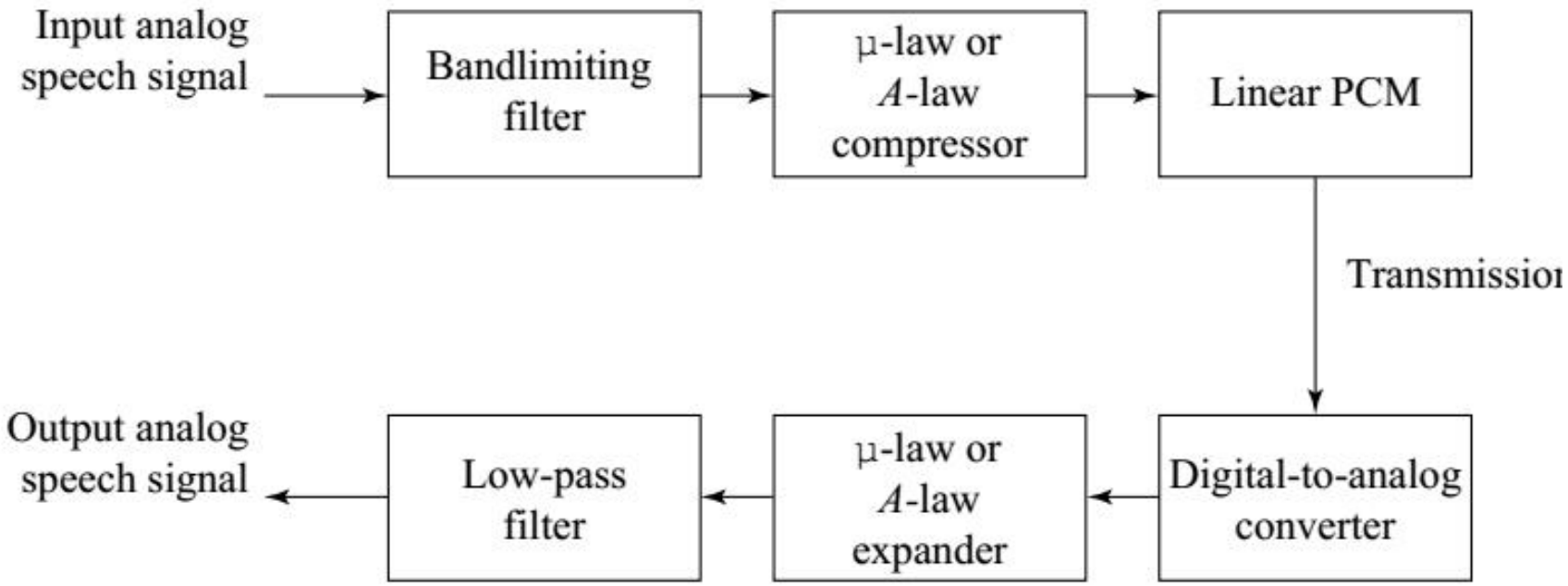


Fig. 6.14: PCM signal encoding and decoding.

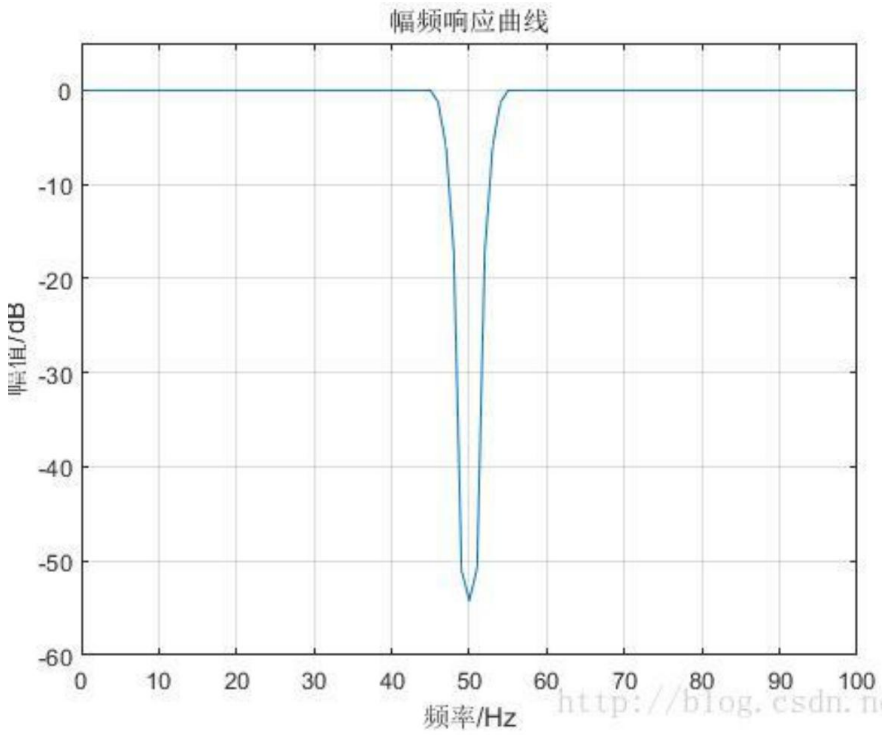
6.1.2 音频滤波

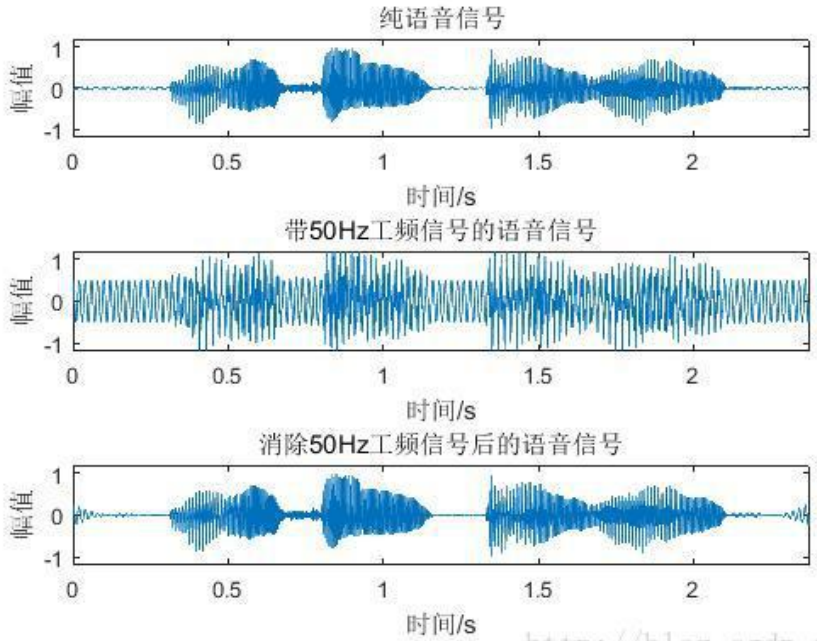
- 对音频信号作滤波消除不需要的频率

- 步骤：
- 1、将音频信号变换到频域空间
- 2、设计滤波器（如50Hz~10kHz的带通滤波器）
- 3、频域操作
- 4、将音频信号变换到时域空间

6.1.2 音频滤波

- 如抑制音频噪声（如工频50Hz）
- 就需要设计一个陷波器。





目 录

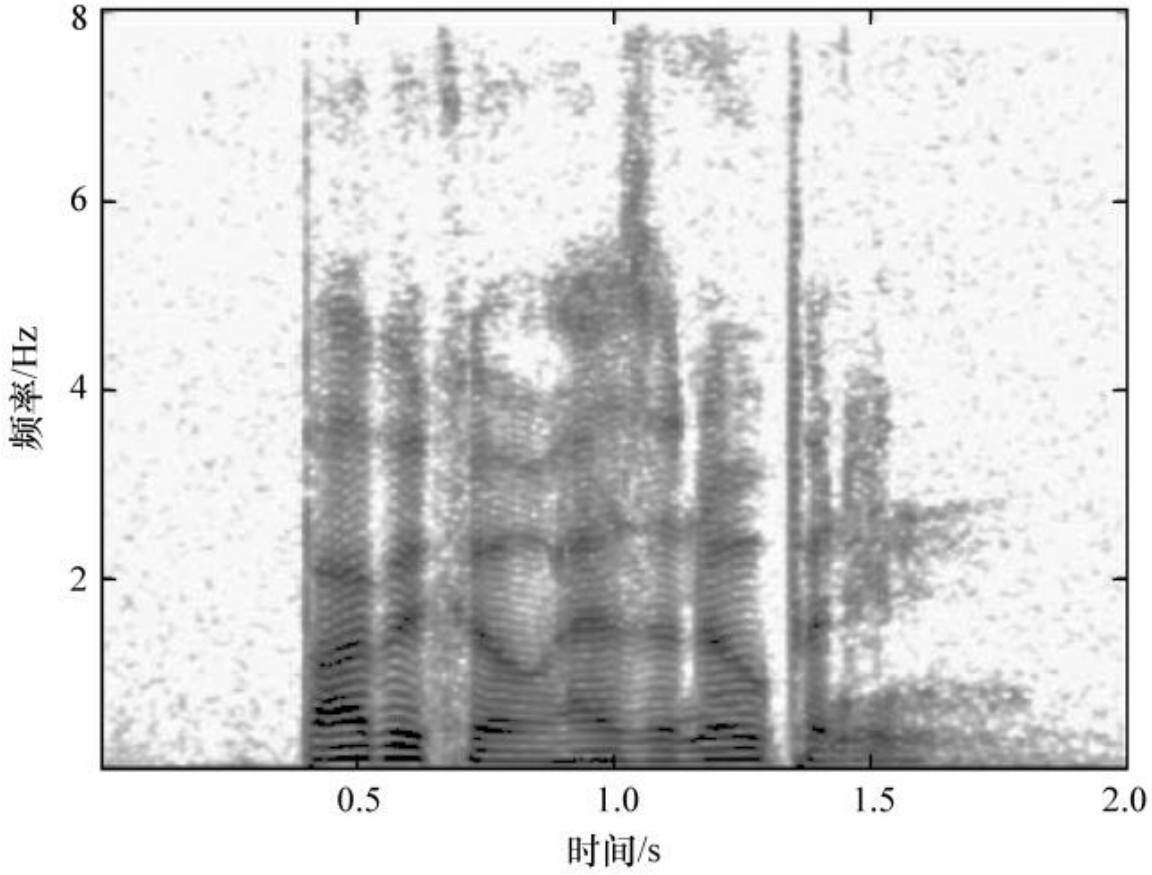
6.2 声学模型

6.2.1 语言的本质

- 语言属人类独有，能让人不费力地交流复杂的思想和感觉
- 组成有声话语的小微语言元素叫作音素，它是语言中最小的单位，一旦改变，单词或者表达就会跟着变化
- 我们的发声器官（舌、下颚、唇）以难以置信的速度和精心的编排在变换着共振结构

6.2.1 语言的本质

- 这是短语“Barbacco has an opening”的语音谱图，横坐标表示时间，纵坐标表示频率。黑色部分表示在一个频率范围内的总能量



6.2.2 声学模型

标准语音系统的组成元素

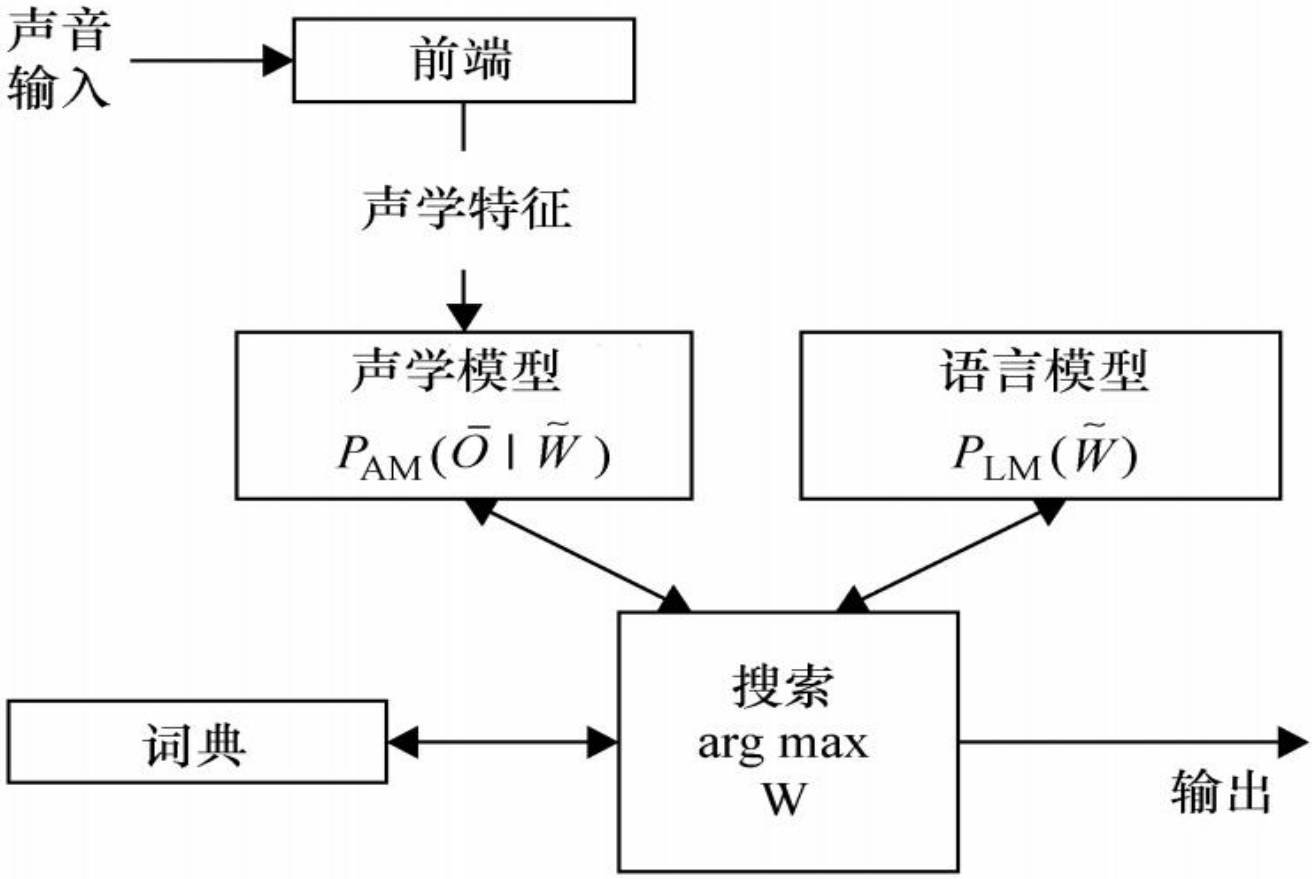


图 3.3　标准语音系统的组成元素

6.2.2 声学模型

标准语音系统

能利用标准语音识别器解决的问题都符合贝叶斯规则（Bayes' rule）：

$$W^* = \arg\max_{\tilde{W}}(P(\tilde{W}|\bar{O})) \tag{3.1}$$

语音识别的目标是找到词组序列的最可能概率 W^*，假设声学观测集 $\bar{O}$，运用贝叶斯规则，我们可以得到：

$$P(\tilde{W}|\bar{O}) = \frac{P(\bar{O}|\tilde{W})P(\tilde{W})}{P(\bar{O})} \tag{3.2}$$

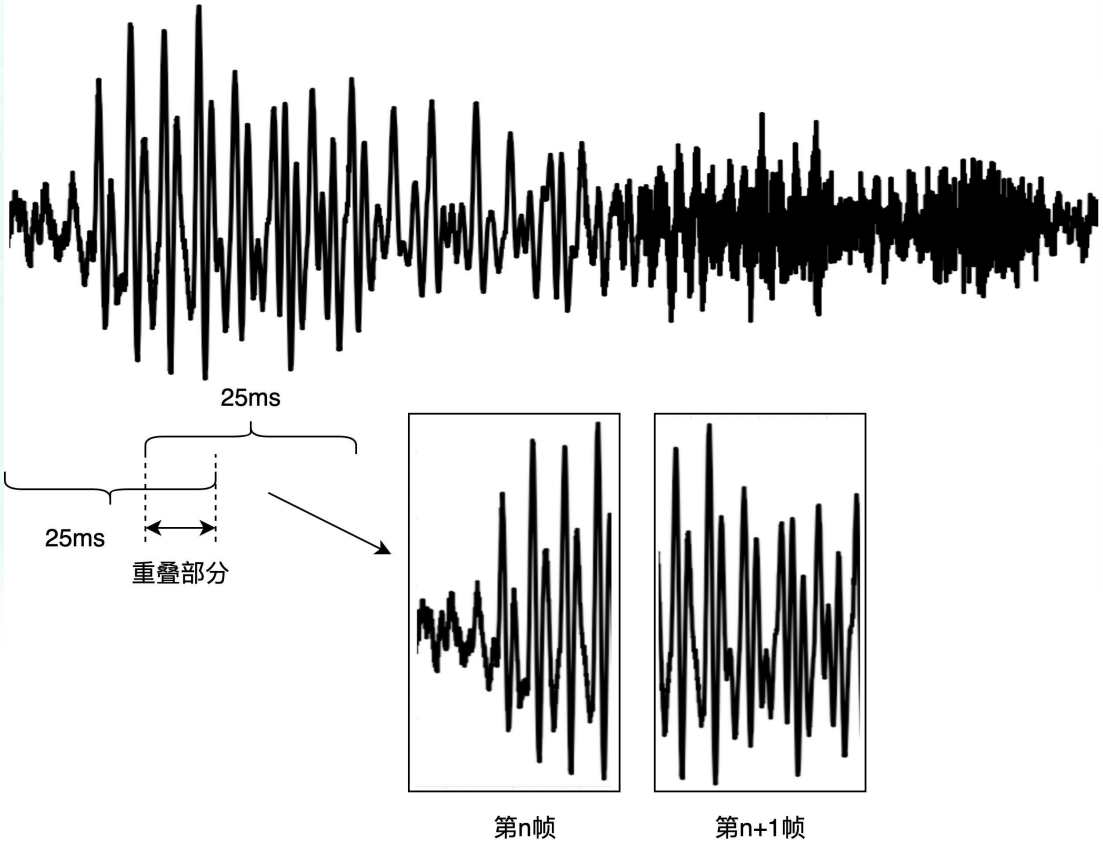
注意到 $P(\bar{O})$ 和词组序列 $\tilde{W}$ 无关，因此我们想要找到：

$$W^* = \arg\max_{\tilde{W}}(P(\bar{O}|\tilde{W})P(\tilde{W})) \tag{3.3}$$

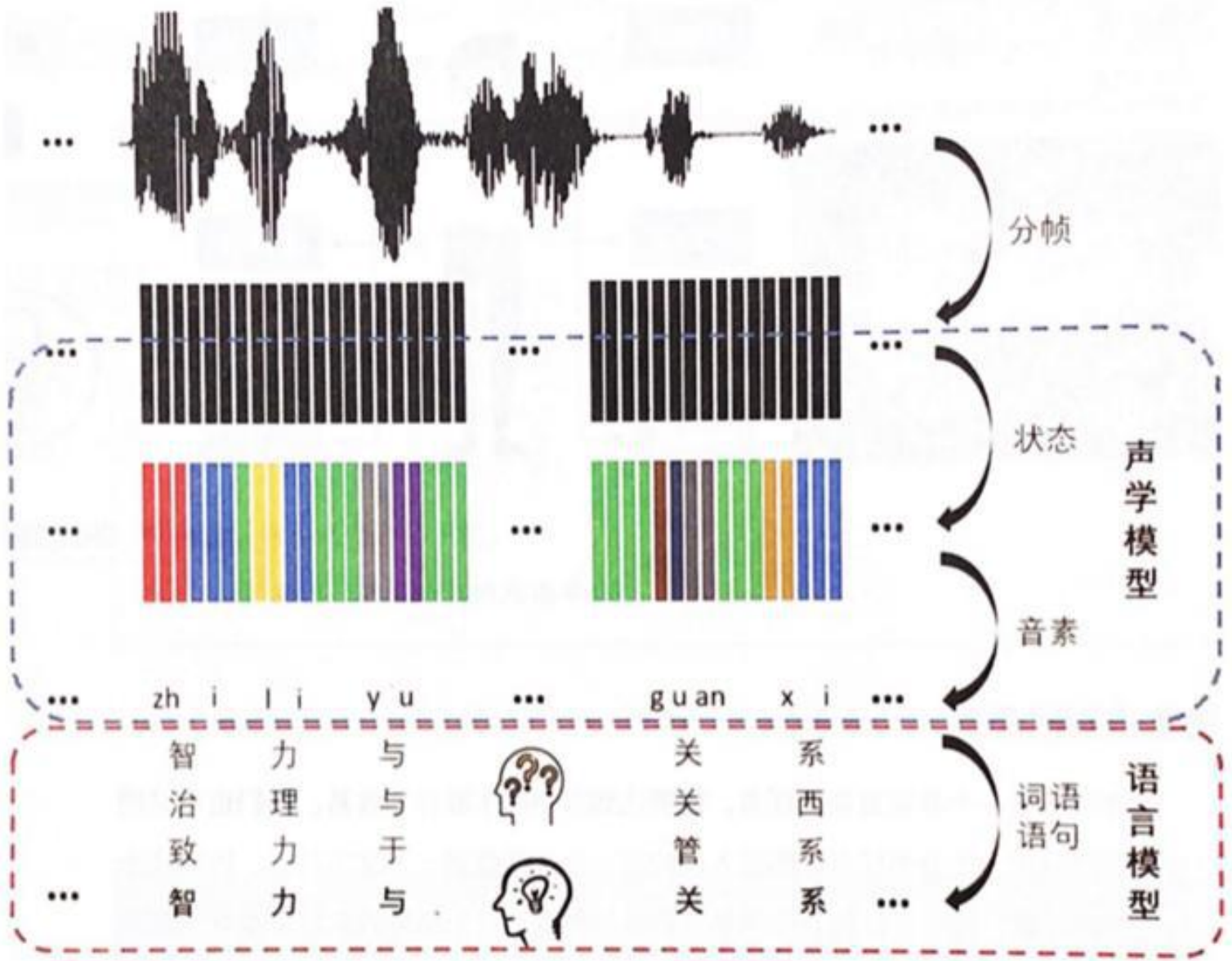
我们使用声学模型（AM）评估 $P(\bar{O}|\tilde{W})$，并用语言模型（LM）评估 $P(\tilde{W})$。

声学模型

- 什么是分帧？通俗地理解就是，加窗处理、分段处理。随着窗口的往右（假设向右代表时间向前）推移，对加窗后的信号逐步展开处理。



6.2.2 声学模型


分帧
状态
音素
声学模型
zh i l i y u g uan x i
智 力 与 关 系
治 理 与 关 西
致 力 于 管 系
智 力 与 关 系
词语
语句
语言模型

6.2.2 声学模型

前端模式

- 前端模式：输入的语言被数字化，并转化成一个矢量序列，它可以找到由一个声学前端输入的整体频谱。

- 声学模型：在一个标准系统里，语言被建模成词组序列，词组则是音素序列。但是声学表达是协同发音的结果，声音和词组里的每一个音素都相互依赖

6.2.2 声学模型

如何计算声学模型

- 怎么才能知道每个单词应该发什么音呢？这就需要另一个模块，叫作词典（lexicon），它的作用就是把单词串转换成音素串。词典一般认为是跟声学模型、语言模型并列的模块。

- 如：词典文件（your_db.dict）一行一个单词，后面空格后跟着的是发音
  - HELLO HH AH L OW
  - WORLD W AO R L D

6.2.2 声学模型

如何计算声学模型

- 有了词典的帮助，声学模型就知道给定的文字串该依次发哪些音了。不过，为了计算语音与音素串的匹配程度，还需要知道每个音素的起止时间。

6.2.2 声学模型

音素与语音信号的匹配

- 声学模型都需要知道怎样计算一个音素与一段语音信号的匹配程度。要做这件事，需要找到一种合适的表示语音信号的方法。一般是把语音信号分成许多帧，对于每一帧，通过傅里叶变换等一系列操作，把它转换成一个特征向量。

6.2.3 音频合成

- 语音合成一般会经过文本与韵律分析、声学处理与声音合成三个步骤，分别依赖于文本与韵律分析模型、声学模型与声码器。其中文本与韵律分析模型一般被称为“前端”，声学模型和声码器被称为“后端”。

6.2.3 音频合成

- 文本与韵律分析中，首先对文本进行分词和标注：分词会将文本切成一个个词语，标注则会注明每个字的发音以及哪里是重音、哪里需要停顿等

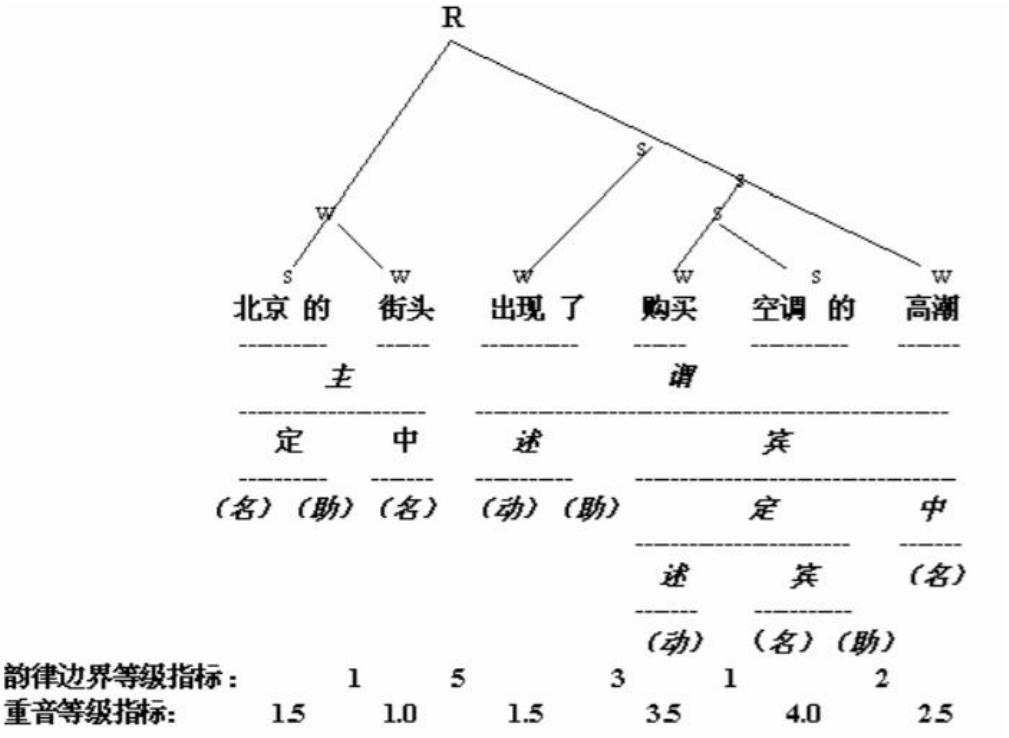


图 3　无标记韵律边界和重音等级预测示例

6.2.3 音频合成

- 数字化信号须转换成模拟信号，才能播放
- 调频方法（参数化合成）
- 波形表法（拼接法）
- AI音频合成

6.2.3 音频合成

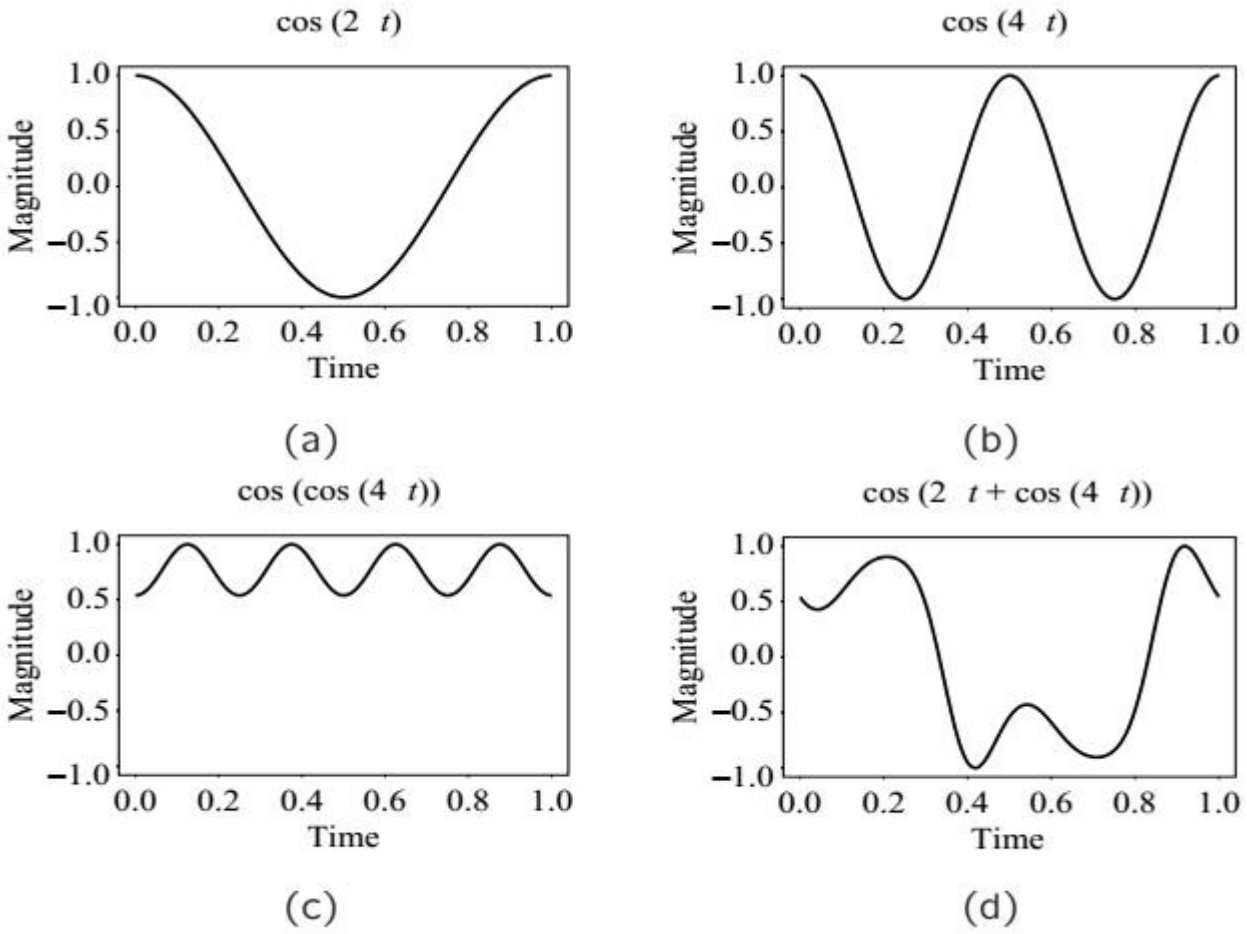
调频方法

1. **FM** (Frequency Modulation): one approach to generating synthetic sound:

$$x(t) = A(t)\cos[\omega_c \pi t + I(t)\cos(\omega_m \pi t + \phi_m) + \phi_c] \qquad (6.11)$$

6.2.3 音频合成

调频方法



6.2.3 音频合成

波形表方法

2. **Wave Table synthesis**: A more accurate way of generating sounds from digital signals. Also known, simply, as **sampling**.

   In this technique, the actual digital samples of sounds from real instruments are stored. Since wave tables are stored in memory on the sound card, they can be manipulated by software so that sounds can be combined, edited, and enhanced.

6.2.3 音频合成

- 如谷歌下一代语音合成系统WaveNet

- Wavenet模型是一种序列生成模型，可以用于语音生成建模。在语音合成的声学模型建模中，Wavenet可以直接学习到采样值序列的映射，因此具有很好的合成效果。目前wavenet在语音合成声学模型建模，vocoder方面都有应用，在语音合成领域有很大的潜力。

- https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

6.2.3 音频合成

MIDI

6.2.3 音频合成

TTS

- http://speech.diotek.com/en/text-to-speech-demonstration.php