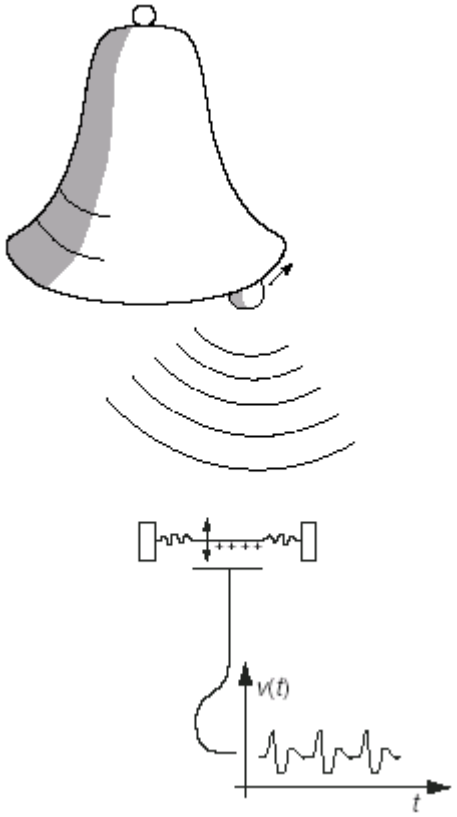
语音信号基础知识

洪青阳 副教授

厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

声音的产生与获取



振动

↓

在空气中形成压力波动

↓

传感器的动作

↓

时变的电压信号

声波的物理描述

- 声波是一种**纵波**，它的振动方向和传播方向是一致的。
- 声波的**频率**是指在单位时间内声波的周期数。声音的频率与声音的**音高**有关。
- 在声学测量中，直接测量声强较为困难，故常用**声压**来衡量声音的强弱。某一瞬间介质中的压强相对于无声波时压强的改变量称为声压，记为p(t)，单位是Pa。

声波的物理描述

- 由于人耳感知的声压动态范围太大，加之人耳对声音大小的感觉近似地与声压、声强呈对数关系，所以通常用**对数值**来度量声音，分别称为声压级与声强级。

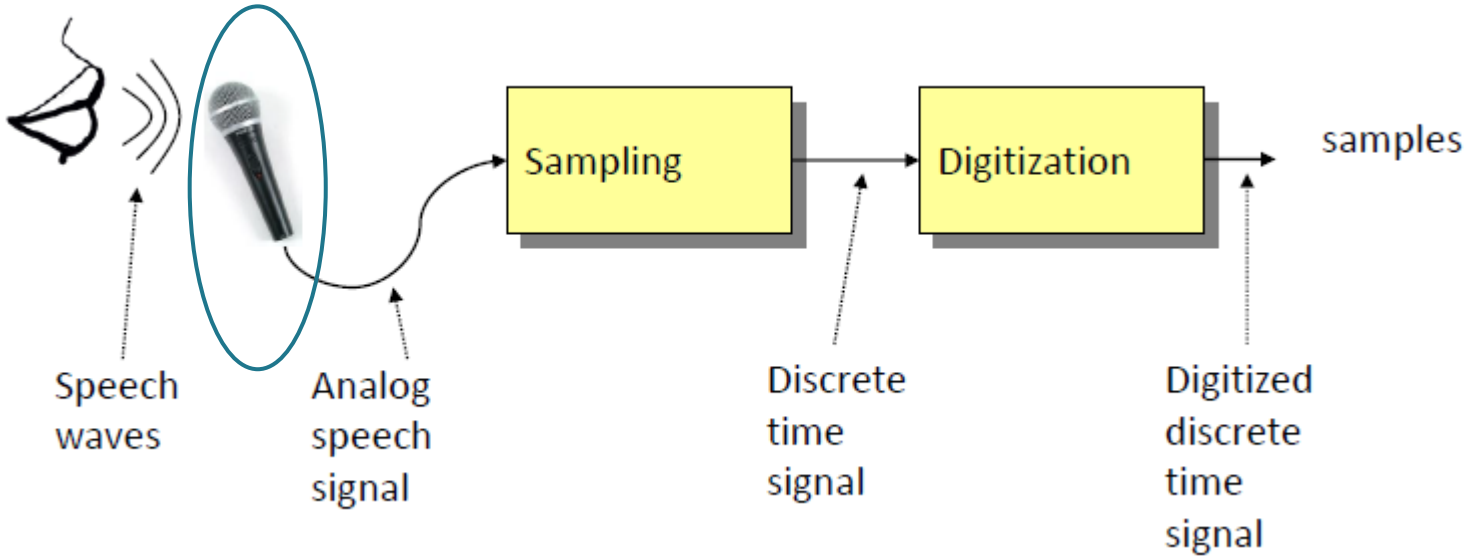
- 一般把很小的声压$p_0=2\times10^{-5}$Pa作为参考声压，把所要测量的声压p与参考声压p_0的比值取常用对数后，乘以20得到的数值称为**声压级**(单位：分贝，dB)

$$\text{声压级} = 20log\left(\frac{p}{p_0}\right)dB$$

安静家庭：35dB，吵闹街道：70dB.

- 按照国家标准规定，住宅区的噪音，白天不能超过50分贝，夜间应低于45分贝，若超过这个标准，便会对人体产生危害。

声音的接收



Thanks to Dr. Ming Li for the contribution of the slides

接收装置的类型

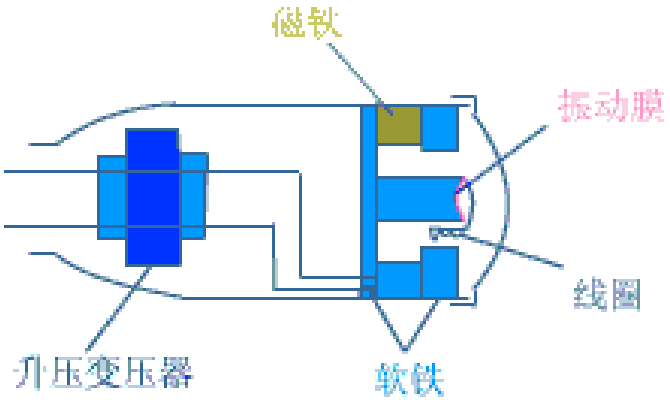


图1 动圈式传声器

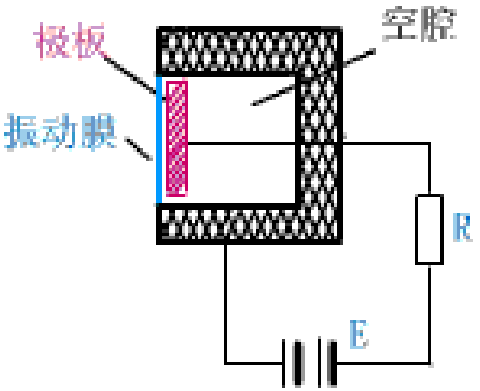
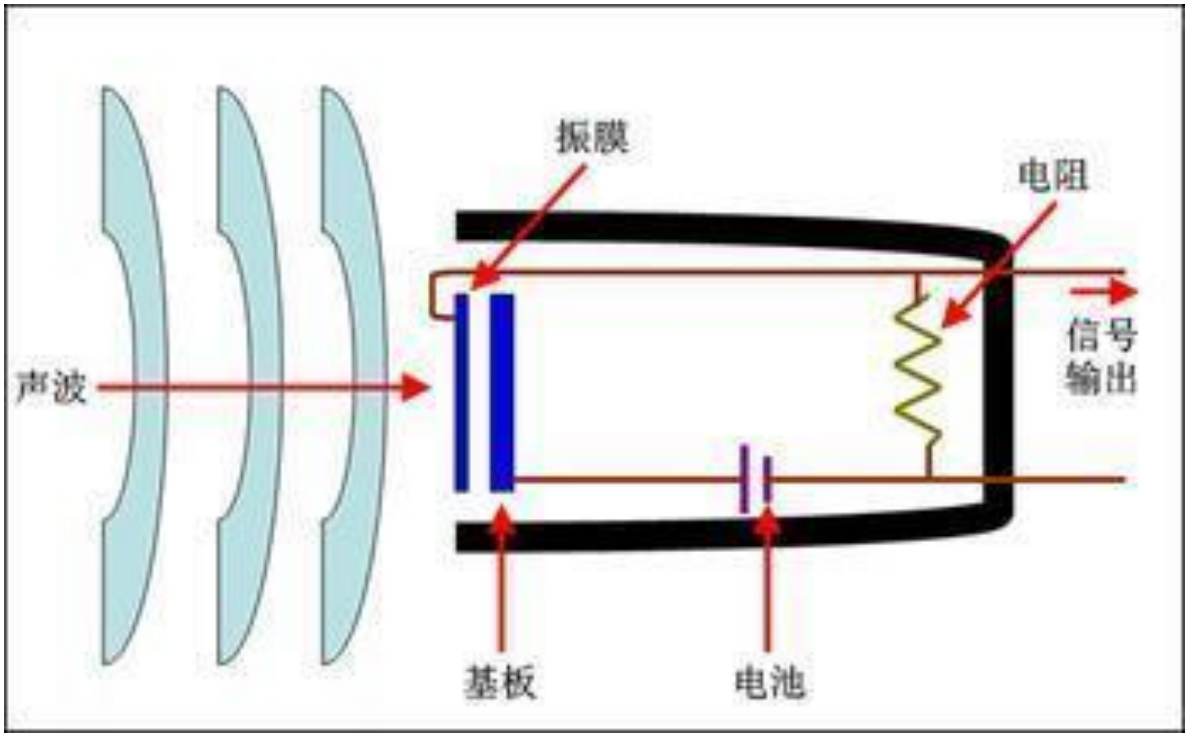


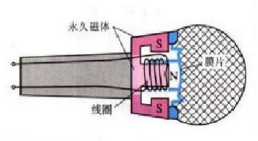
图2 普通电容式传声器

声音的接收原理

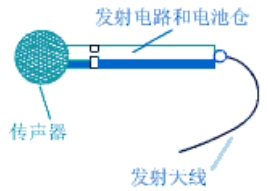


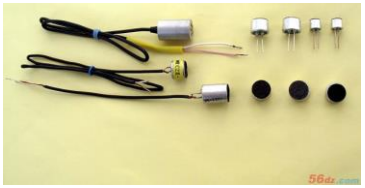
声音的接收装置

- 麦克风(又称传声器、拾音器、话筒)的类型
  - **动圈式麦克风**：精度、灵敏度较低，体积大。其突出特点是输出阻抗小，所以接较长的电缆也不降低其灵敏度。温度和湿度的变化对其灵敏度也无大的影响。用于语音广播、扩声系统。

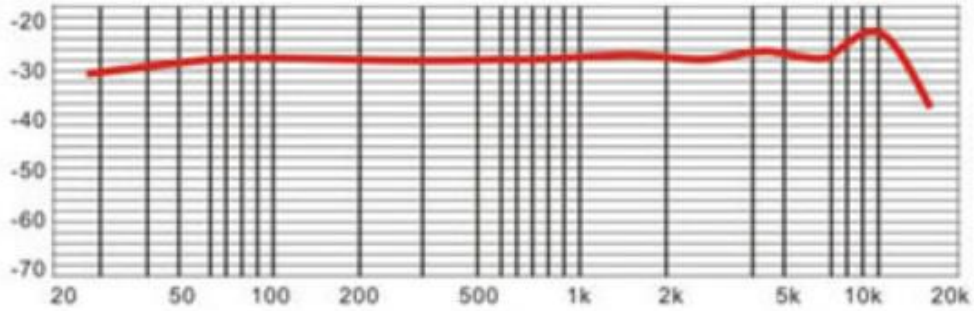


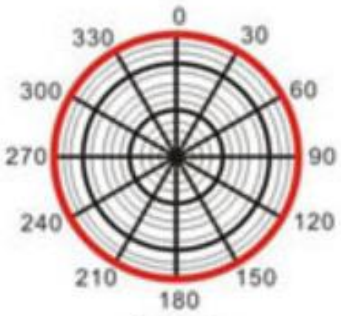
  - **电容式麦克风**：音质好，灵敏度较高，但需要电源，用于舞台、录音室等。
    - 驻极体麦克风：是电容式的一种，无需外加电源，体积小，使用最广泛。
      - **振膜式**：极化带电体是驻极体振膜本身，驻极体话筒拾声的音质效果相对差些，多用在对于音质效果要求不高的场合，如普通电话机、玩具等。
      - **背级式**：极化带电体是涂敷在背极板上的驻极体膜层，将储存电荷的膜层与振膜分离设计，手机、语音识别等高端传声录音产品多采用背极式驻极体。





频 响 图





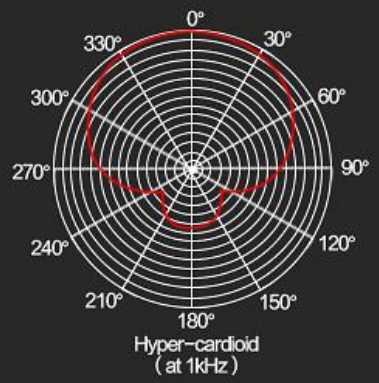
全指向
(at 1kHz)

性能指标

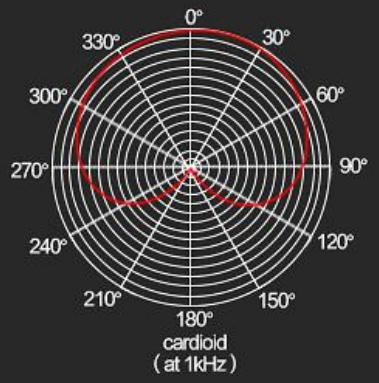
- 类型：电容式（驻极体）
- 指向性：全指向
- 频率响应：20HZ-16kHZ
- 灵敏度：-30dB±2dB RL=2.2kΩ VS=3.0V
- 插头类型：3.5MM标准插头（直插型）
- 线长：1.4-5米可选
- 重量：30g
- 包装：透明袋装

SGC-578技术参数

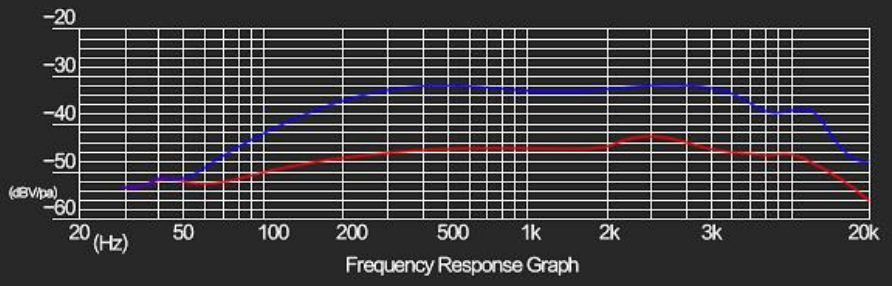
- 单体：背极式驻极体
- 指向性：心型指向/超心型指向
- 频率响应：80Hz-14kHz
- 灵敏度：-30dB±2dB (0dB=1V/Pa at 1kHz)
- 输出阻抗：500Ω / 1600Ω ±30% (at 1kHz)
- 负载阻抗：≥1000Ω
- 供电方式：1.5V AA电池
- 单体尺寸：Ø22 x 278mm



Hyper-cardioid
(at 1kHz)



cardioid
(at 1kHz)



Frequency Response Graph

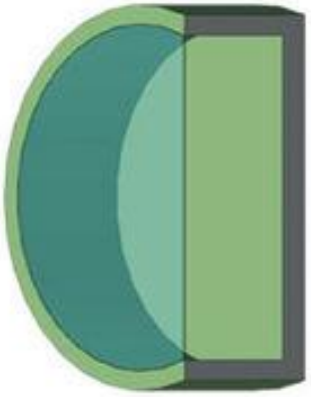
Hyper-cardioid　Cardioid

麦克风的性能指标

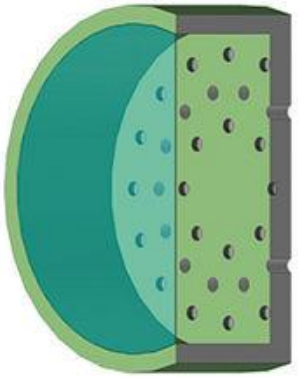
- **灵敏度**：在一定强度的声音作用下输出电信号的大小。以分贝表示，并规定1V/Pa为0dB，因话筒输出一般为毫伏级，所以，其灵敏度的分贝值始终为负值。

指向性

- 话筒对于不同方向来的声音灵敏度会有所不同，这称为话筒的方向性。
- 方向性用传声器正面0方向和背面180方向上的灵敏度的差值来表示，差值大于15dB者称为强方向性话筒。



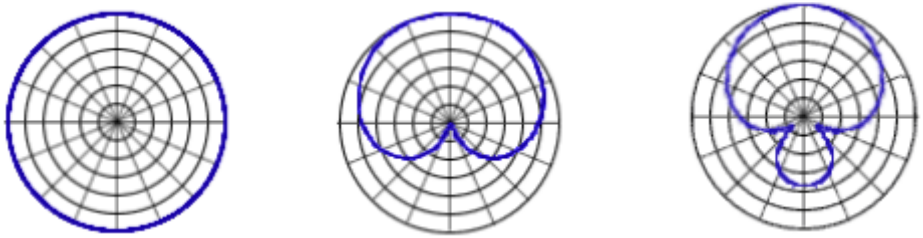
全指向性



单一指向性

指向性

- 全指向性话筒从各个方向拾取声音的性能一致。当说话者要来回走动时采用此类话筒较为合适，但在环境噪声大的条件下不宜采用。
- 心形指向话筒的灵敏度在水平方向呈心脏形，正面灵敏度最大侧面稍小，背面最小。这种话筒在多种扩音系统中都有优秀的表现。
- 单指向性话筒又称为超心形指向性话筒，它的指向性比心形话筒更尖锐，正面灵敏度极高，其它方向灵敏度急剧衰减，特别适用于高噪音的环境。

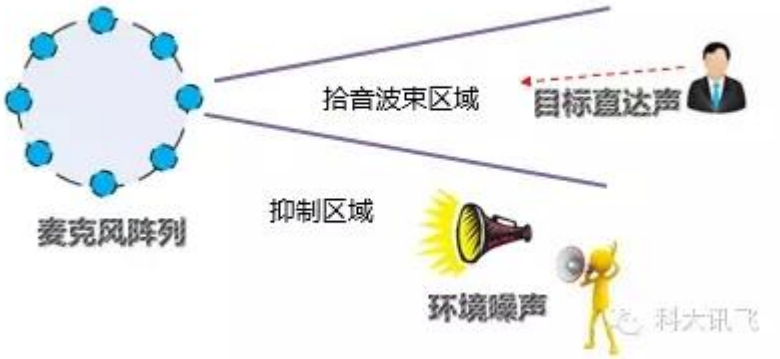


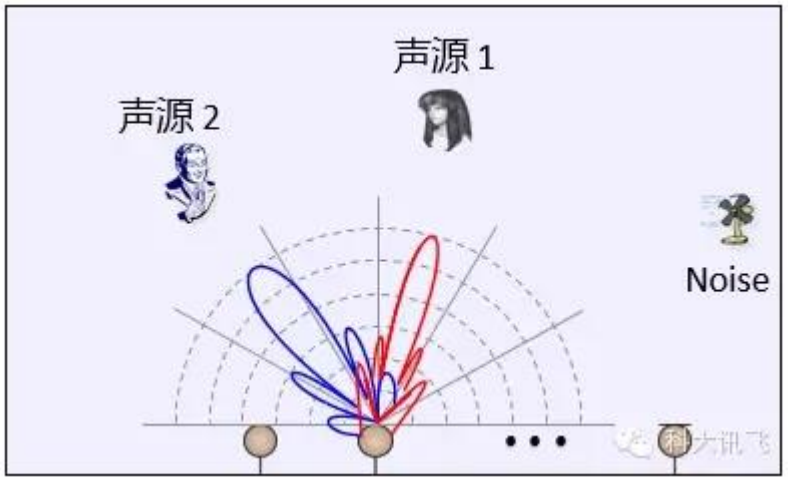
输出阻抗

- 目前常见的话筒有高阻抗与低阻抗之分。高阻抗的输出电压略高，但引线电容所起的旁路作用较大，使高频下降，同时也易受外界的电磁场干扰，所以，话筒引线不宜太长，一般以10～20米为宜。
- 低阻抗输出无此缺陷，所以噪音水平较低，传声器引线可相应的加长，有的扩音设备所带的低阻抗传声器引线可达100米。如果距离更长，就应加前级放大器。

麦克风阵列

① 语音增强(Speech Enhancement)
② 声源定位(Source Localization)
③ 去混响(Dereverberation)
④ 声源信号提取(分离)





麦阵远场声源分离

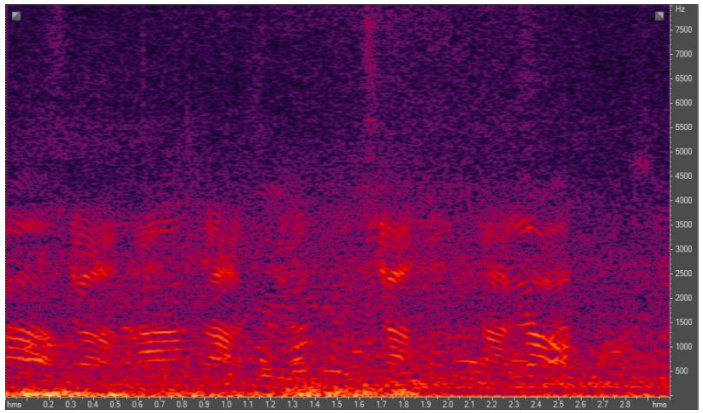
10m距离测试, 室外条件下不同方向两个说话人同时说话

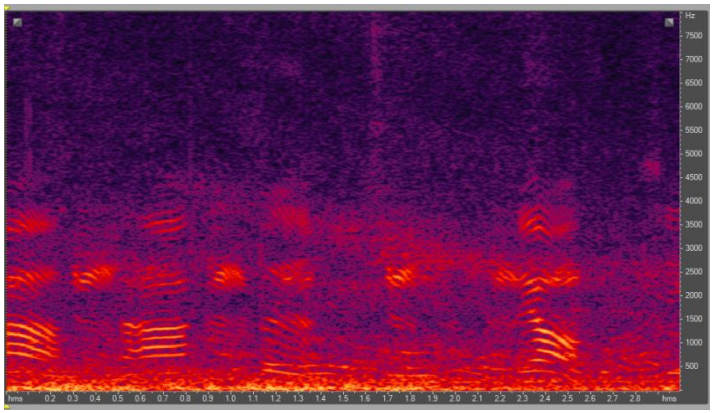
“厦门大学”

“1,2,3,4,5,6”

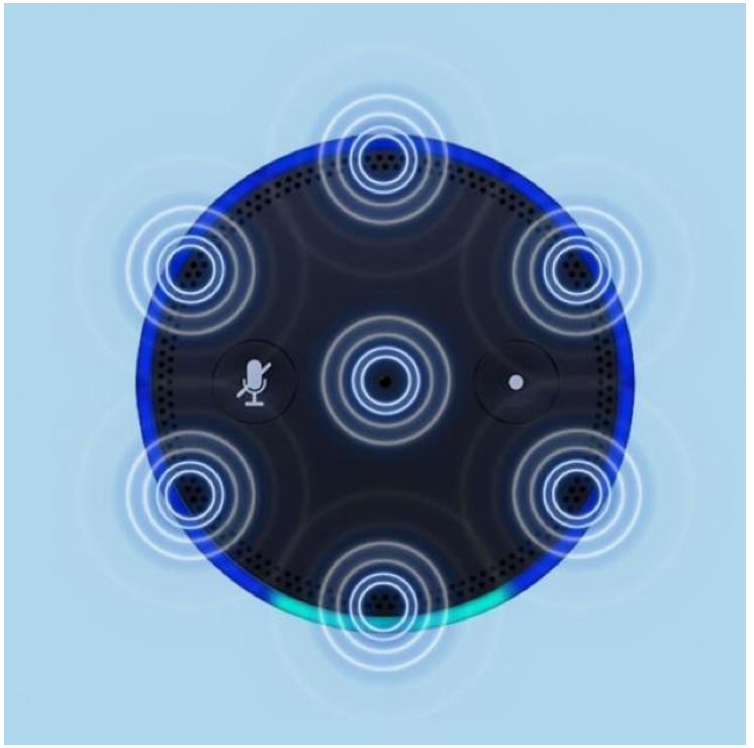
原始语音

麦阵对准右边说话人输出

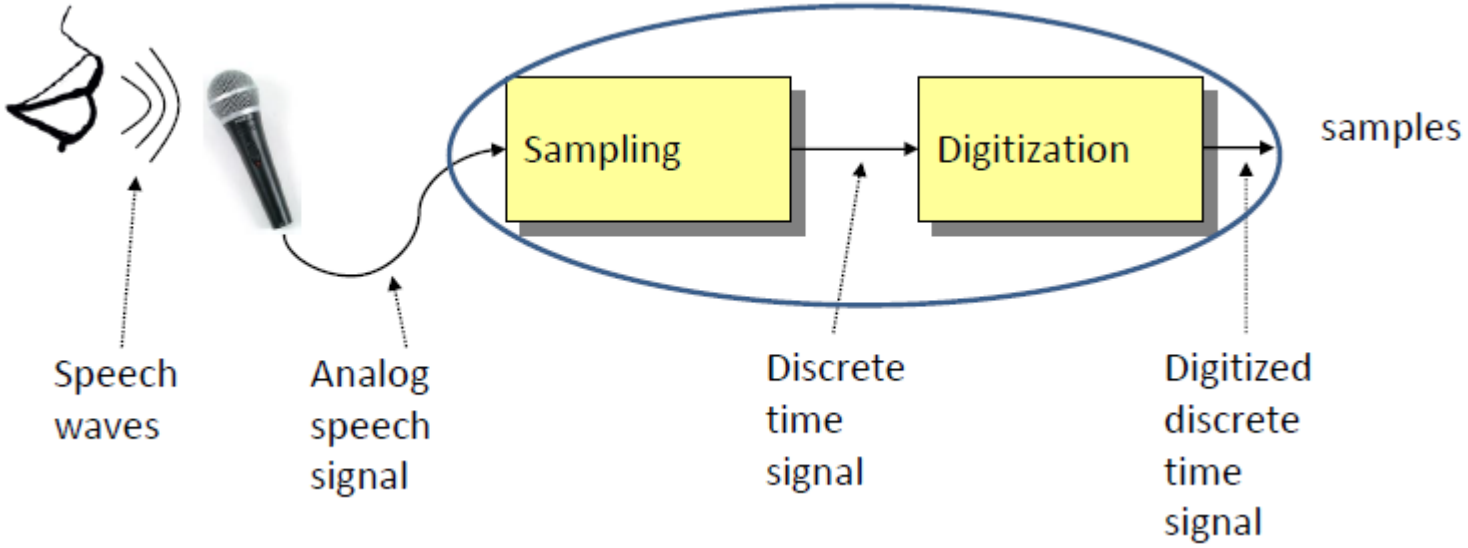




亚马逊Echo采用6+1麦克风阵列

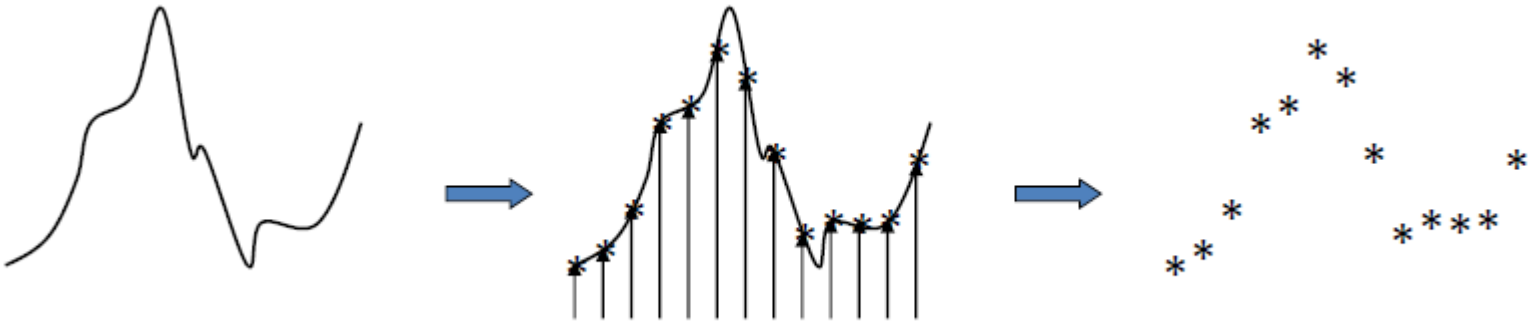


采样和量化



Thanks to Dr. Ming Li for the contribution of the slides

采样(Sampling)

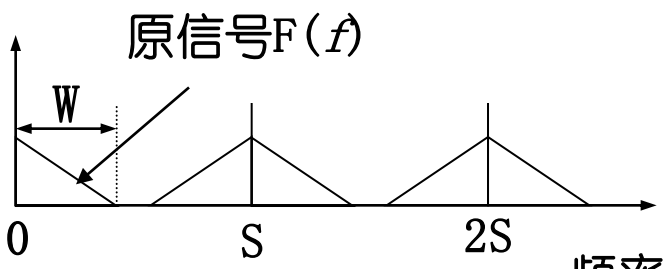


采样标准：能够重现声音，与原始语音尽量一致
采样率：每秒采样点数。

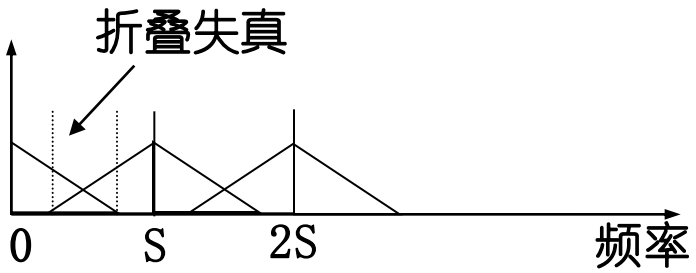
Thanks to Dr. Ming Li for the contribution of the slides

采样定理

- 当采样率大于信号中最高频率的2倍时，采样之后的数字信号完整地保留了原始信号中的信息。采样定理又称奈奎斯特(Nyquist)定理。

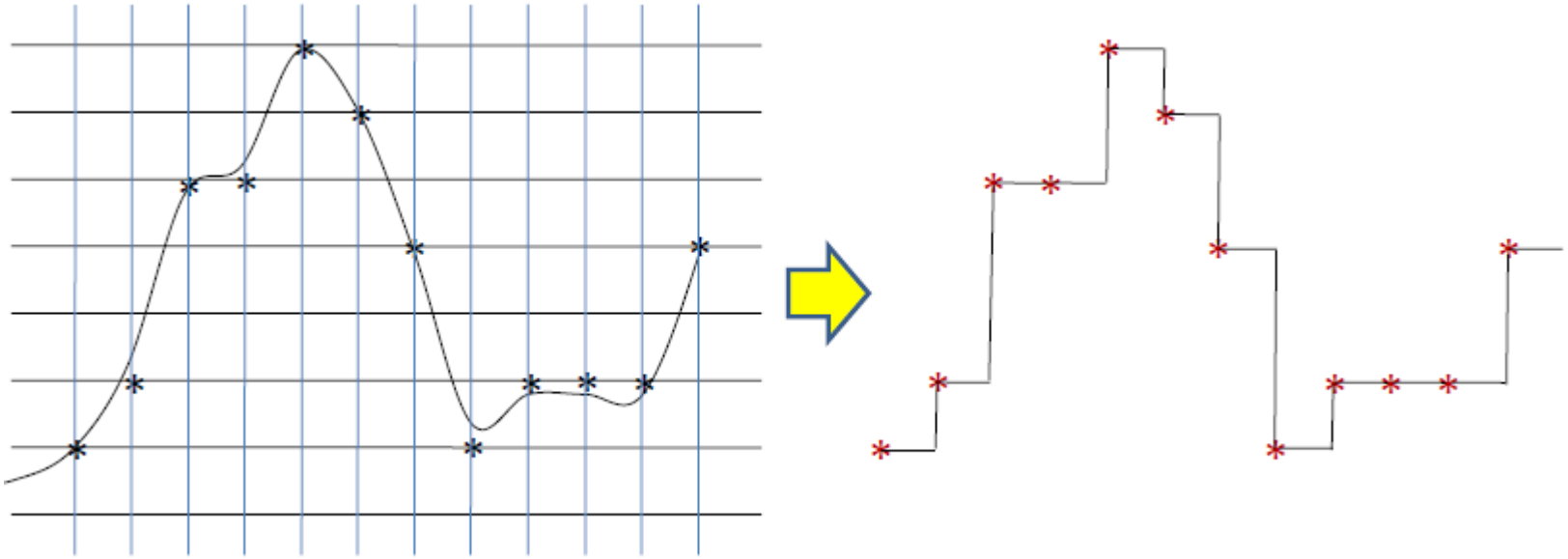


(a)准确取样(S≥2W)



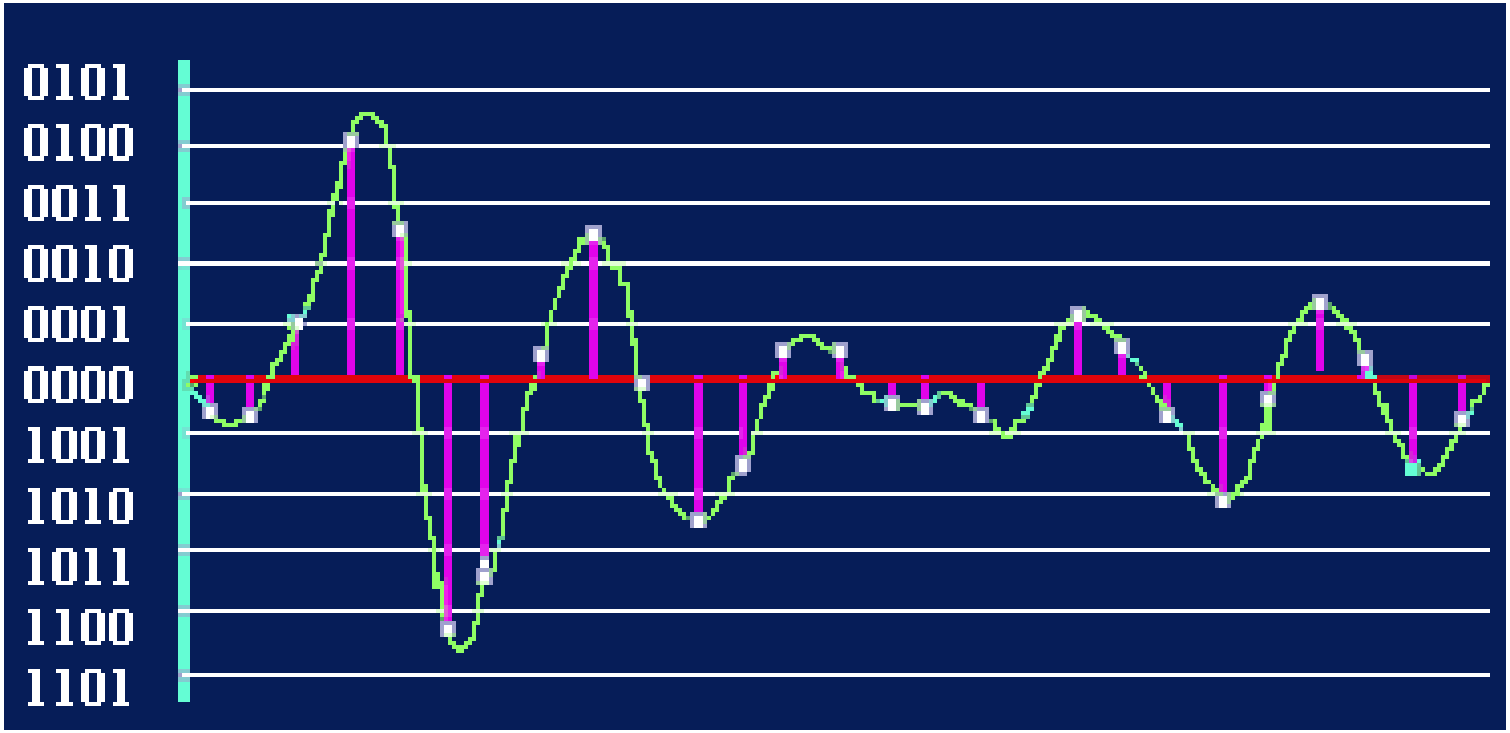
(b) 非准确取样时(S＜2W)

量化(Digitization)



Thanks to Dr. Ming Li for the contribution of the slides

量化过程



量化：采样点的值不能取任意值
取值范围：2^N-1，N=8, 16, 32
必须为整型。

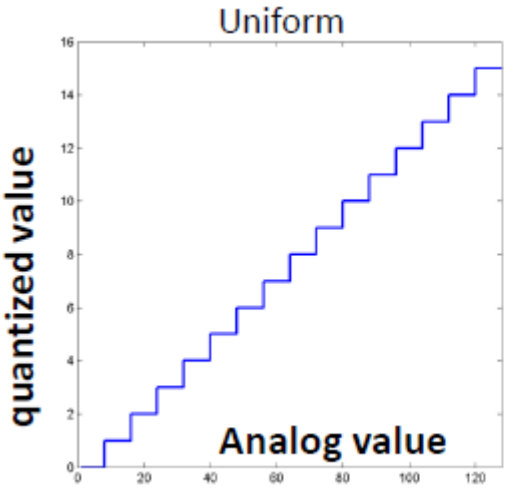
量化过程

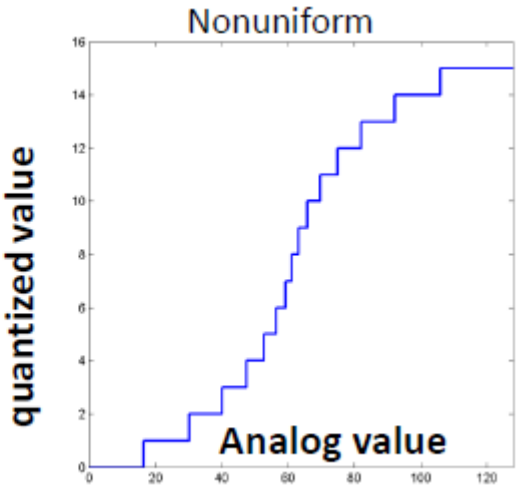
模拟电压、量化和编码		
电压范围(V)	量化(十进制数)	编码(二进制数)
0.5～0.7	3	0011
0.3～0.5	2	0010
0.1～0.3	1	0001
-0.1～0.1	0	0000

- 量化过程是指将每个采样值在幅度上再进行离散化处理。

量化过程

- 量化方法
  - 均匀量化
  - 非均匀量化



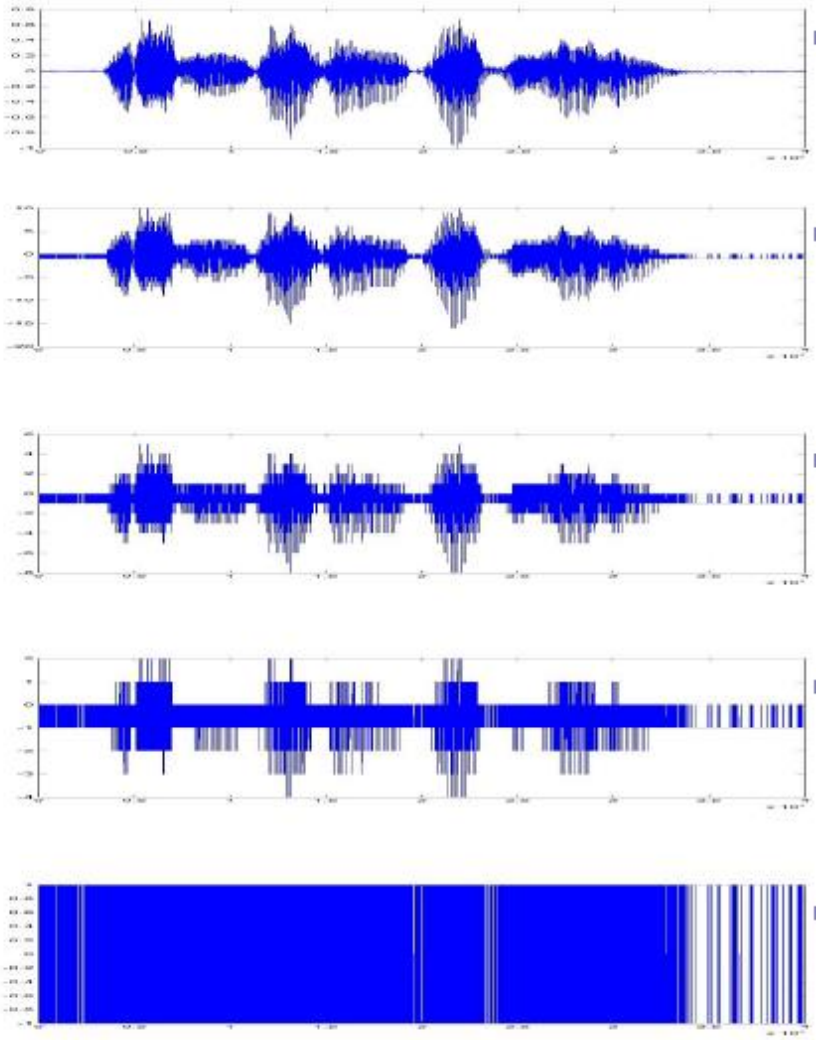


- 量化误差
  - 量化会引入失真，并且量化失真是一种不可逆失真，这就是通常所说的量化噪声。
  - 信噪比(signal-to-noise ratio，SNR)

$$SNR(dB) = 10\lg(\frac{\sigma_x^2}{\sigma_e^2}) = 6.02B + 4.77 - 20\lg(\frac{X_{\max}}{\sigma_x})$$

量化失真

- 16 bit sampling
- 5 bit sampling
- 4 bit sampling
- 3 bit sampling
- 1 bit sampling



Thanks to Dr. Ming Li for the contribution of the slides

语音文件格式的重要参数

- **采样率**：8K(电话、嵌入式), 16K(PC), 44.1K(CD)
- **采样精度(量化位数)**：即每次取样信息量。
- **码率**（bps: bits per second，又称比特率），如8k16bit为128kbps.
- **语音通道数**：语音通道的个数表明语音产生的波形数，一般分为单声道和立体声道。单声道产生一个波形，立体声道则产生两个波形。

常用的音频编码格式

- PCM编码
  1. 脉冲编码调制(pulse code modulation,PCM)是将模拟信号经采样、量化、编码的过程。它只将编码后的数据保存，并不保存任何格式信息。最大优点是音质好，最大缺点是体积大。
  2. PC麦克风常用格式(宽带录音,16k16bit)。可保存为PCM raw data(.raw文件, 无头部)或Microsoft PCM格式(.wav文件)。
  3. ADPCM编码是有损编码(32kbps)，保存为Microsoft ADPCM格式(.wav文件) 。

  存储格式：
  –PCM raw data (*.raw)
  –Microsoft PCM (*.wav)
  –Microsoft ADPCM (*.wav)

常用的音频编码格式

- MP3

  MP3对音频信号采用的是有损压缩方式，压缩率高达10:1~12:1。为了降低声音失真度，MP3采取了“感官编码技术”，并使压缩后的文件回放时能够达到比较接近原始音频数据的声音效果。

常用的音频编码格式

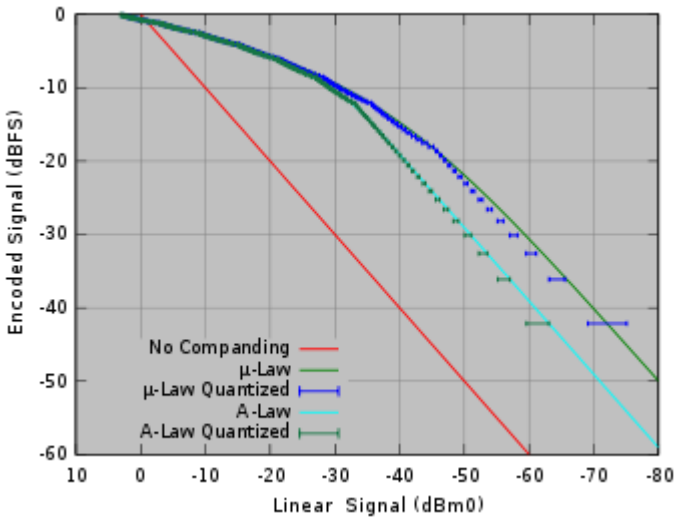
- NIST

  包含ASCII文本格式的1024字节可读的Sphere头部。以下示例使用16位线性PCM和16KHz的采样率。

```
NIST_1A
    1024
sample_rate -i 16000
channel_count -i 1
sample_coding -s3 pcm
sample_count -i 65380
sample_n_bytes -i 2
sample_byte_format -s1 01
language -s7 English
handset -s12 SamsungNexus
database_id -s7 RSR2015
end_head
```

常用的音频编码格式

- A-law(A律)编码
  1. ITU-T（国际电联电信标准局）定义的关于脉冲编码的一种压缩/解压缩算法。
  2. 世界上大部分国家采用A律压缩算法。美国采用mu律算法进行脉冲编码。
  3. 固话录音(300-3300Hz)常用的格式(窄带录音，8k8bit)。

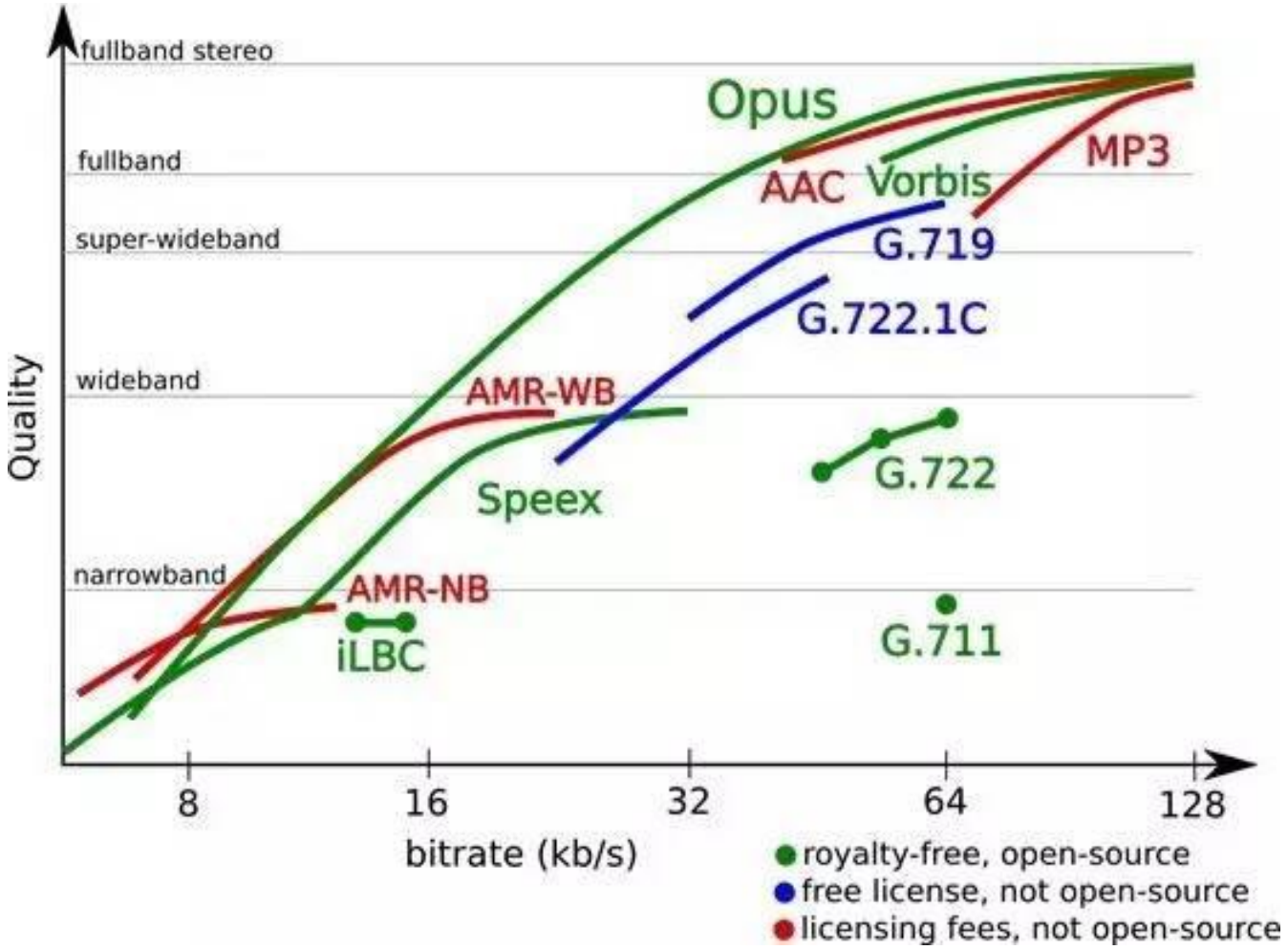


https://en.wikipedia.org/wiki/A-law_algorithm

语音文件常见格式

- WAV (Microsoft PCM): PC麦克风录音常用格式。
- A-law(8k8bit): 电话录音常用格式。
- MP3
- AMR(Adaptive Multi-Rate):每秒钟的AMR音频大小可控制在1K左右，常用于彩信、微信语音，但失真比较厉害。
- WMA(Windows Media Audio): 微软公司推出的与MP3格式齐名的一种新的音频格式，在压缩比和音质方面都超过了MP3。
- AAC(Advanced Audio Coding): 相对于MP3，AAC格式的音质更佳，文件更小。
- M4A: MPEG-4 音频标准的文件的扩展名，最常用的.m4a文件是使用AAC格式的。
- FLAC(Free Lossless Audio Codec): FLAC是一套著名的自由音频压缩编码，其特点是无损压缩。
- NIST(*.sph)：语音识别常用格式，包含ASCII文本格式的1024字节可读的头。

音频编码标准


fullband stereo
fullband
super-wideband
wideband
narrowband
Quality
Opus
AAC
Vorbis
MP3
G.719
G.722.1C
AMR-WB
Speex
G.722
AMR-NB
iLBC
G.711
8
16
32
64
128
bitrate (kb/s)
royalty-free, open-source
free license, not open-source
licensing fees, not open-source

音频编码标准

	算　法	名　称	码率/(kb/s)	标准	制定组织	制定时间	应用领域	质量
波形编码	PCM (A/μ)	压扩法	64	G.711	ITU	1972	PSTN ISDN	4.3
	ADPCM	自适应差值量化	32	G.721	ITU	1984		4.1
	SB-ADPCM	子带 ADPCM	64/56/48	G.722	ITU	1988		4.5
参数编码	LPC	线性预测编码	2.4		NSA	1982	保密语音	2.5
混合编码	CELPC	码激励 LPC	4.8		NSA	1989		3.2
	VSELPC	矢量和激励 LPC	8	GIA	CTIA	1989	移动通信语音信箱	3.8
	RPE-LTP	长时预测规则码激励	13.2	GSM	GSM	1983		3.8
	LD-CELP	低延时码激励 LPC	16	G.728	ITU	1992	ISDN	4.1
	MPEG	多子带感知编码	128	MPEG	ISO	1992	CD	5.0

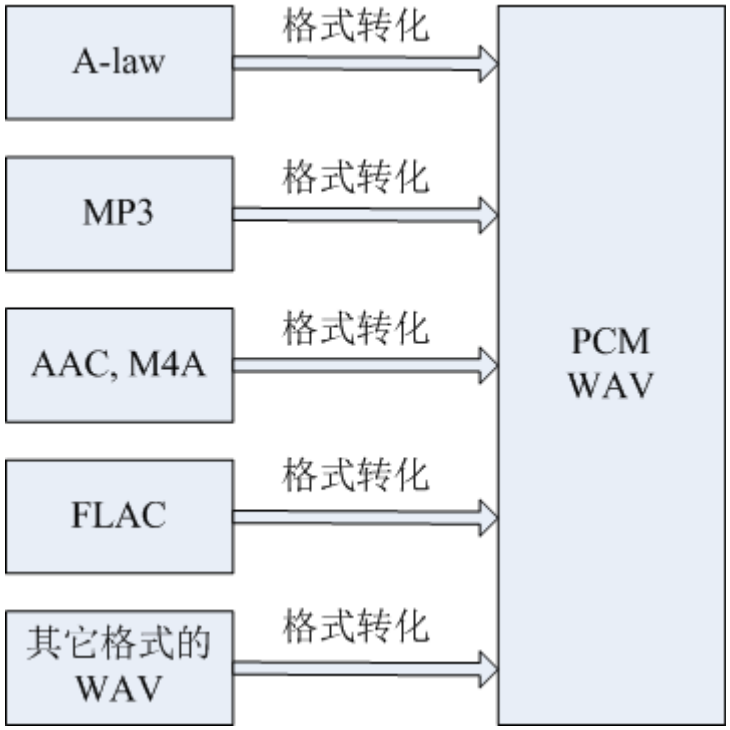
Speex编码

编解码算法	码率范围（kbps）	压缩等级	压缩率
speex-wb	3.95~42.2	0~10	5.98~58.18
speex	2.15~24.6	0~10	5.08~45.71

Speex编解码库压缩率变换范围较广，压缩等级可供选择的范围较宽，所以应用在网络状况较为复杂的移动终端应用中甚为合适。 ——科大讯飞

格式转化

- 基于PCM编码的WAV常作为不同编码互相转换时的一种中介格式，以便于后续处理。



音频工具：
Cool Edit
Adobe Audition
格式工厂

例：8bit A-law转16bit PCM

```
#define SIGN_BIT        (0x80)/* Sign bit for a A-law byte. */
#define QUANT_MASK      (0xf)   /* Quantization field mask. */
#define NSEGS           (8)     /* Number of A-law segments. */
#define SEG_SHIFT       (4)     /* Left shift for segment number. */
#define SEG_MASK        (0x70)/* Segment field mask. */

short  alaw2linear2(unsigned char a_val)  {
    short t;
    short seg;

    a_val ^= 0x55;
    t = (a_val & QUANT_MASK) << 4;
    seg = ((unsigned short)a_val & SEG_MASK) >> SEG_SHIFT;
    switch (seg) {
    case 0:
        t += 8;
        break;
    case 1:
        t += 0x108;
        break;
    default:
        t += 0x108;
        t <<= seg - 1;
    }
    return ((a_val & SIGN_BIT) ? t : -t);
}
```

WAV格式的技术组成

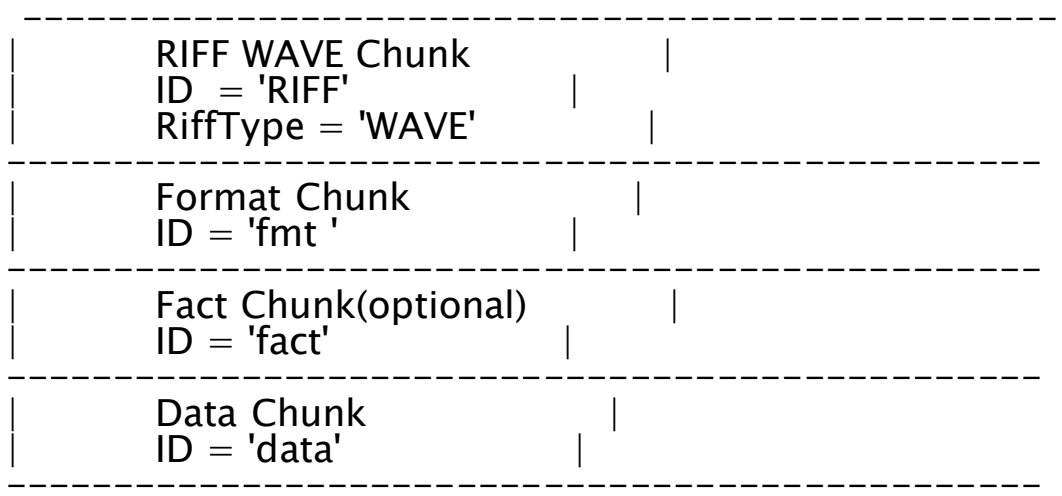
在音频信息处理中，经常需要编程处理不同格式的音频数据，但前面介绍的音频格式多数不公开源码，因此也很难将它们一一详细介绍，这里只介绍WAV格式的技术构成。

WAV

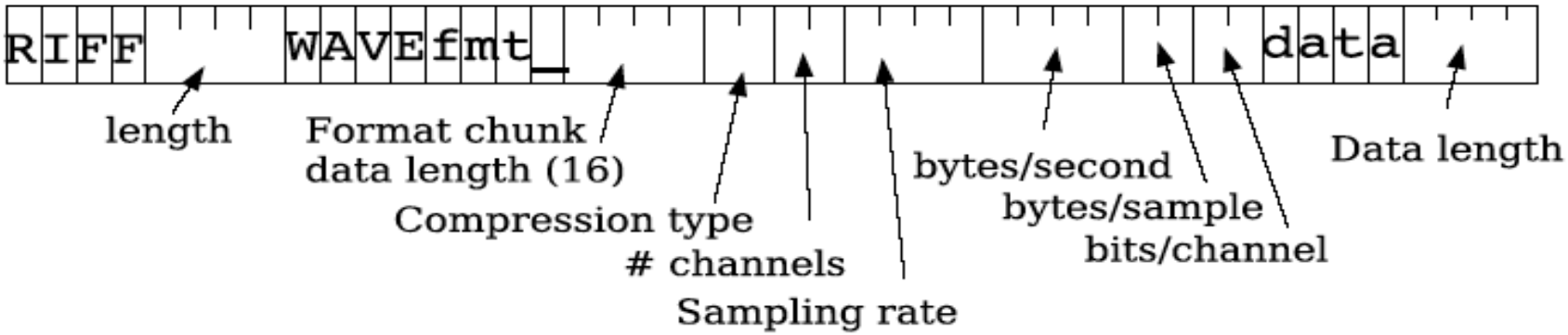
- 计算机中最常见的存放声音格式，就是WAV文件格式，其扩展名是 .wav。
- WAV文件是以RIFF（resource interchange file format）的档案格式储存，含有不定长度的文件头(header)与数据(data)，组成不定长度的区块(chunk)与子区块(sub-chunks)，所存的数据是编码的声音信号，WAV文件支持线性波形编码调制(PCM)、自适应差分脉冲编码调制(ADPCM)等波形编码实现。

WAV

- WAVE文件是由若干个Chunk组成的。按照在文件中的出现位置包括：RIFF WAVE Chunk, Format Chunk, Fact Chunk(可选), Data Chunk。具体见下图：



WAV format


RIFF
WAVEfmt_
data
length
Format chunk
data length (16)
Compression type
channels
Sampling rate
bytes/second
bytes/sample
bits/channel
Data length

WAV头部(PCM文件)

```
typedef struct {
    char riff[4];          // RIFF file identification (4 bytes)
    long length;           // length field (4 bytes)
    char wave[4];          // WAVE chunk identification (4 bytes)
}WAVECHUNK;

typedef struct{
    char fmt[4];           // format sub-chunk identification  (4 bytes)
    long flength;          // length of format sub-chunk (4 byte integer)
    short format;          // format specifier (2 byte integer)
    short chans;           // number of channels (2 byte integer)
    long sampsRate;        // sample rate in Hz (4 byte integer)
    long bpsec;            // bytes per second (4 byte integer)
    short bpsample;        // bytes per sample (2 byte integer)
    short bpchan;          // bits per channel (2 byte integer)
}FMTCHUNK;

typedef struct{
    char data[4];          // data sub-chunk identification  (4 bytes)
    long dlength;          // length of data sub-chunk (4 byte integer)
}DATACHUNK;
```

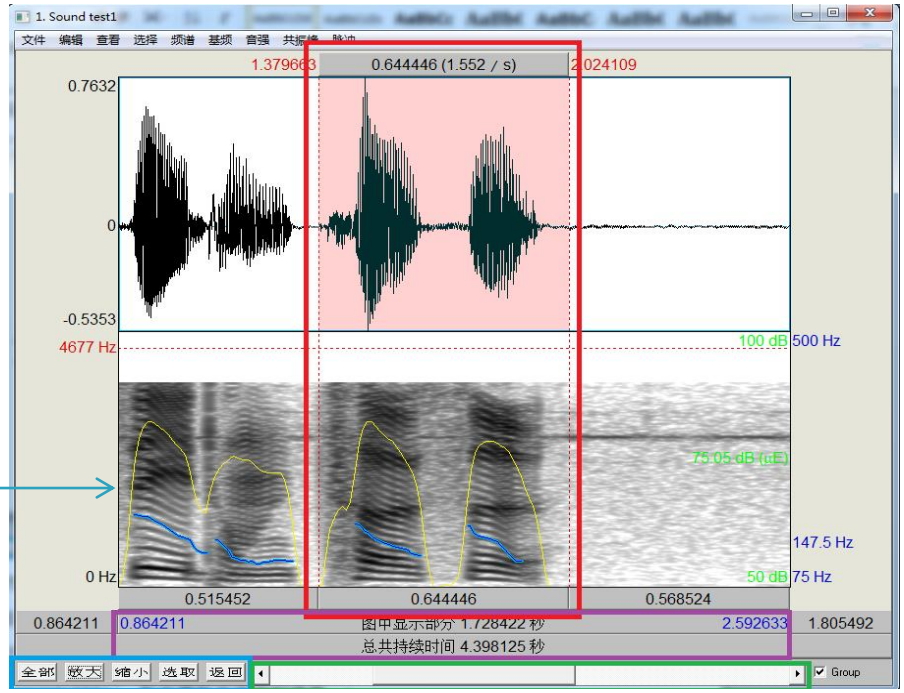
WAVE文件格式说明表

	偏移地址	字节数	数据类型	内　容
文件头	00H	4	char	"RIFF"标志
	04H	4	long	文件长度
	08H	4	char	"WAVE"标志
	0CH	4	char	"fmt"标志
	10H	4		过渡字节（不定）
	14H	2	short	格式类别（10H为PCM形式的声音数据)
	16H	2	short	通道数，单声道为1，双声道为2
	18H	4	long	采样率（每秒样本数），表示每个通道的播放速度，
	1CH	4	long	波形音频数据传送速率，其值为通道数×每秒数据位数×每样本的数据位数 / 8。播放软件利用此值可以估计缓冲区的大小。
	20H	2	short	数据块的调整数（按字节算的），其值为通道数×每样本的数据位值 / 8。播放软件需要一次处理多个该值大小的字节数据，以便将其值用于缓冲区的调整。
	22H	2	short	每样本的数据位数，表示每个声道中各个样本的数据位数。如果有多个声道，对每个声道而言，样本大小都一样。
	24H	4	char	数据标记符 " data "
	28H	4	long	语音数据的长度

总共44个字节

语谱图

可以用时间–频域–幅度的方式显示出原始声音的语谱图。



语谱图

总结

- 麦克风类型
- 采样和量化过程
- 语音文件常见格式
- WAV头部
- 语谱图

Thank you!

Any questions?