语音活动性检测方法

李军锋

中国科学院声学研究所

中科院语言声学与内容理解重点实验室

2023年10月12日

提纲

- 简介
- 原理
  - 信号模型
  - 基本假设
- 基于短时声学特征的方法
- 基于能量特征与统计模型的方法
- 基于多维特征与机器学习的方法

提纲

- 简介
- 原理
  - 信号模型
  - 基本假设
- 基于短时声学特征的方法
- 基于能量特征与统计模型的方法
- 基于多维特征与机器学习的方法

语音活动性检测的背景和意义

- 语音识别领域的市场近些年来处于高速增长阶段
- 进一步挖掘识别器的识别能力，国内外在技术上已经遇到了难以突破的瓶颈
- 前端信号处理系统左右着识别系统在实际声学环境中的表现，它仍然有较大的提高余地

语音活动性检测的作用

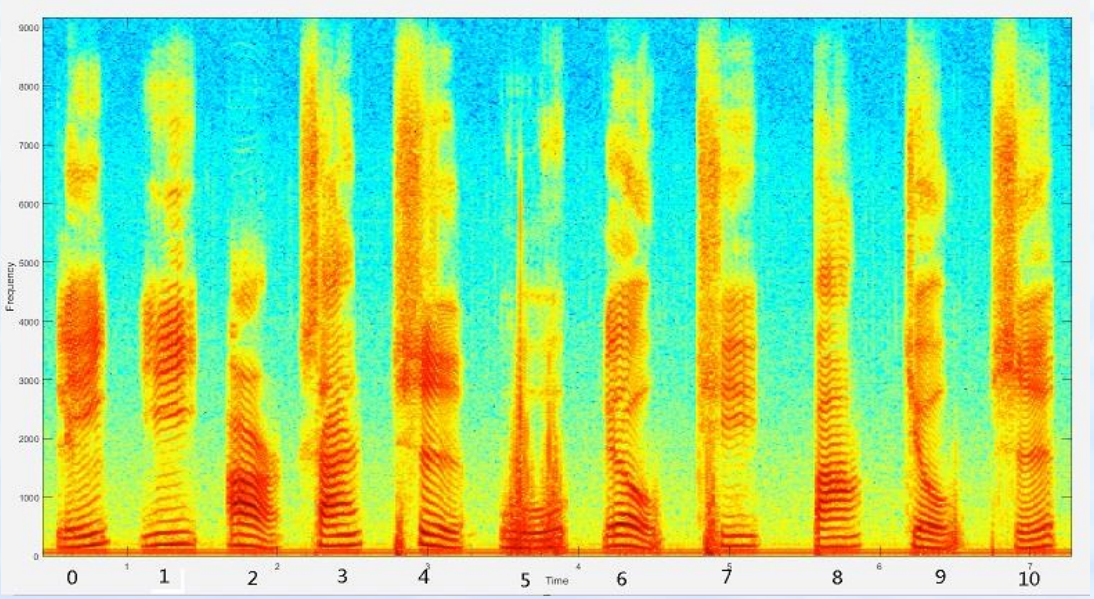
- 在大多数稳健语音识别算法中，需要追踪语音或者噪声的特征，以便对声学模型或者特征作出补偿。
- 如果在识别语音信号之前，将非语音信号从输入信号中剔除，那么识别器的似然概率就能更准确的描述语音信号本身，从而避免非语音部分的干扰。从理论上讲，这种方法能够提高识别器的精度。
- 在连续变化的声学环境中，我们很难建立一个精确的噪声模型来补偿声学模型或者特征，但我们可以提前将非语音信号从输入信号中剔除，从而尽可能的限制非稳定噪声对于识别器的影响。
- 剔除非语音信号，能够降低识别系统的计算量，特别是非语音信号所占的比率较大时，作用非常明显。

语音活动性检测的作用

- **影响用户对识别器的使用体验**
  - ✓ 对非语音信号的误操作发生的频率
  - ✓ 识别系统对输入的语音信号作出响应的时延

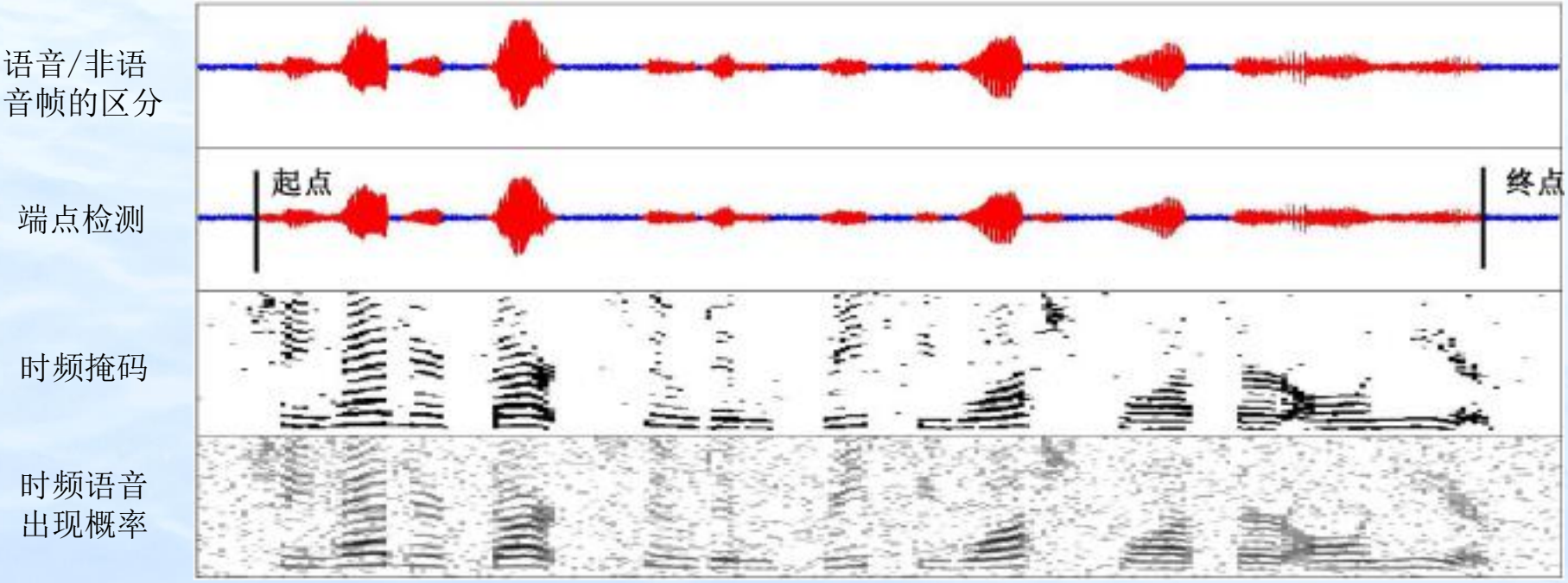
语音信号的特点

- 语音信号的能量主要集中少数诸如谐波结构等特征的时频点上，通常称为语音信号时频域的稀疏分布。
- 语音信号处理主要关注语音信号出现的部分，被噪声信号占据的部分通常用于噪声统计特征的估计。
- 甚至有些应用中（例如语音增强），对于包含语音信号的帧，还需要排除不含语音信号的时频点。



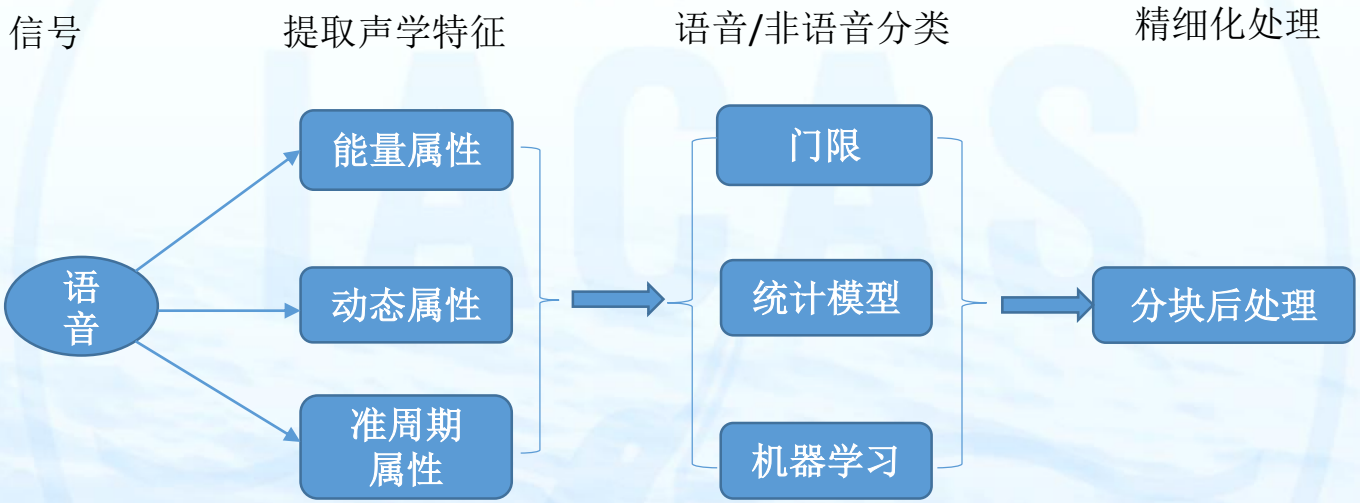
语音活动性的表达方式

- 语音/非语音帧的区分：对各帧信号进行二分类，用于某些语音编码和识别。（Voice Activity Detection, VAD）
- 端点检测：确定句子的时间起始点和终点，忽略中间少量的非语音帧，用于语音识别。(Speech Endpoint Detection)
- 时频掩码：对时频点进行分类，用于听觉感知、语音增强、语音分离等，0表示语音缺失，1表示语音出现。(Speech Mask Estimation)
- 时频语音出现概率：对时频点上的语音出现的概率描述，用途同上



语音端点检测简介

- 语音活动检测的本质--------语音与非语音的二分问题



- 特别说明
  - 这里语音特指包含语音的信号，这部分信号有可能叠加有环境噪声。
  - 非语音信号是指不包含语音的信号，这部分信号有可能是静音、不包含语音的时频点、以及纯噪声。

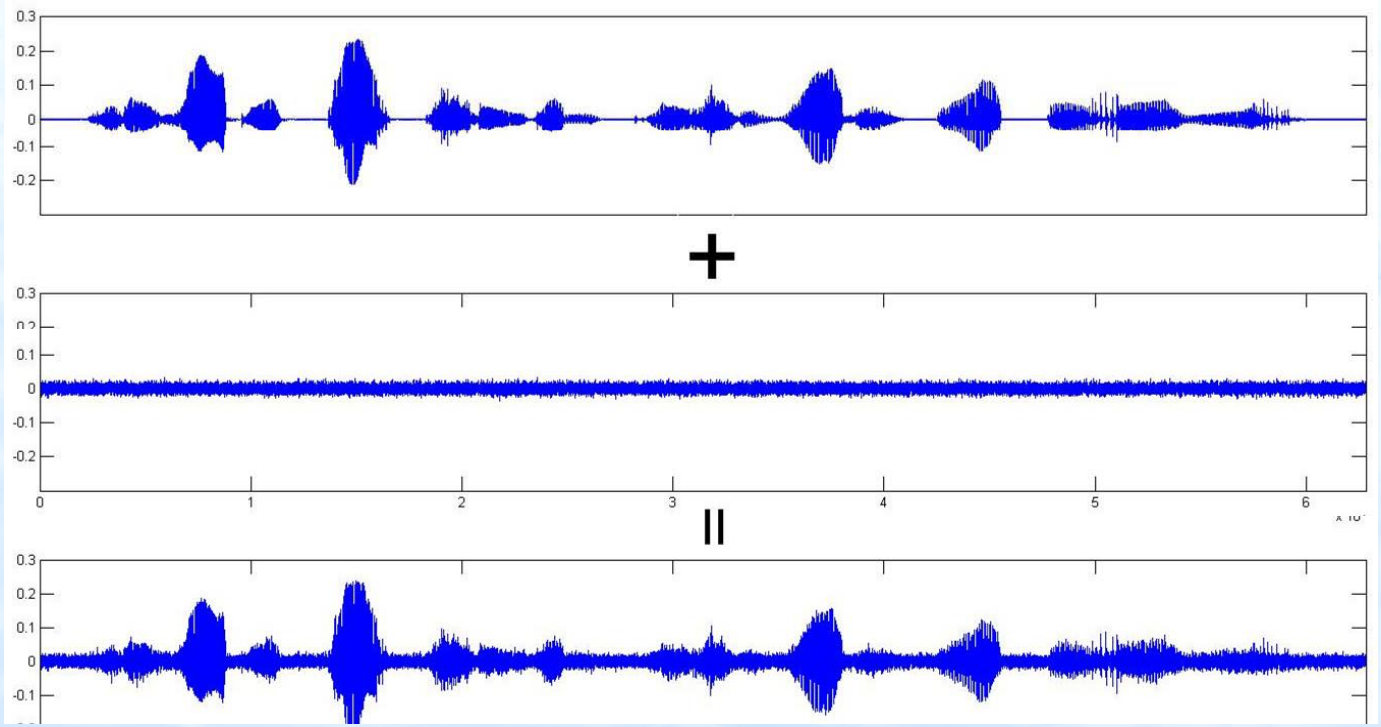
提纲

- 简介
- 原理
  - 信号模型
  - 基本假设
- 基于短时声学特征的方法
- 基于能量特征与统计模型的方法
- 基于多维特征与机器学习的方法

基本原理

- 考虑一段带有加性噪声的语音信号，y(t)表示观察信号的幅度值，s(t)表示纯净语音信号，n(t)表示噪声信号。在时间域，信号模型满足加性关系如下：

$$y(t)=s(t)+n(t)$$



纯净语音

噪声

可观察信号

基本原理

- 对于连续的波形信号，取一帧信号加窗，进行快速傅里叶变换（STFT），在频域信号模型表示为：

$$y(\omega) = s(\omega) + n(\omega)$$

- 对傅里叶系数取幅度平方值，得到功率谱：

$$|y(\omega)|^2 = |s(\omega)|^2 + |n(\omega)|^2 + 2|s(\omega)| \times |n(\omega)| \times \cos\theta$$

这里的θ表示两个虚数$s(\omega)$和$n(\omega)$的相位差值。对方程两边取数学期望得到：

$$\mathrm{E}(|y(\omega)|^2) = \mathrm{E}(|s(\omega)|^2) + \mathrm{E}(|n(\omega)|^2) + 2\mathrm{E}(|s(\omega)| \times |n(\omega)| \times \cos\theta)$$

假设噪声与语音是两个无关的向量，那么$\mathrm{E}(|s(\omega)| \times |n(\omega)| \times \cos\theta)$=0

最终，噪声功率谱的关系可以近似表示为：

$$|y(\omega)|^2 \approx |s(\omega)|^2 + |n(\omega)|^2$$

基本原理

- 通常采用对数值描述功率谱包络，相应地：

$$\log(|y(\omega)|^2) = \log(|s(\omega)|^2 + |n(\omega)|^2) > \log(|n(\omega)|^2)$$

- 在假设噪声信号相对于语音信号稳定的情况下，语音的包络幅度应该大于非语音包络幅度，对数谱包络也同样如此。相应地，语音谱包络的高阶差分值应该大于非语音包络的差分值。
- 能量是在噪声稳定假设的前提下，使用最为广泛的特征。

提纲

- 简介
- 原理
- 基于短时声学特征的方法
  - 谱熵
  - 过零率
  - 能量
  - 谐波结构
- 基于能量特征与统计模型的方法
- 基于多维特征与机器学习的方法

短时幅度谱熵特征

物理意义

- 熵在信息论中是反映信息度量的一个量。某随机事件的随机性越大，即不确定性越高，则熵值也越大，所以携带的信息量亦越大。

计算方法

- 对于一帧信号，进行傅里叶变换提取幅度谱
- 归一化幅度谱，得到类似概率密度函数的表达：

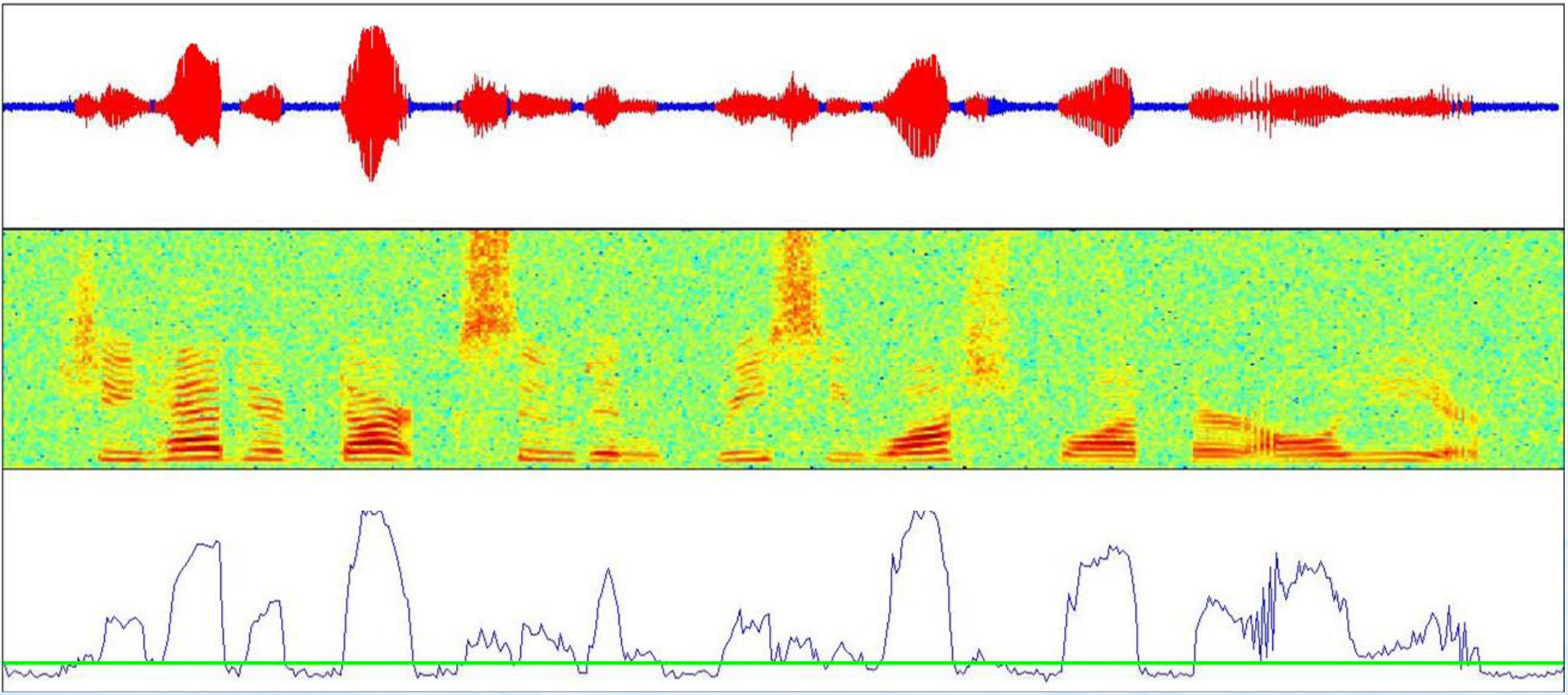
$$p_i = \frac{s(f_i)}{\sum_{k=1}^{N} s(f_k)}, i = 1, \ldots, N$$

- 计算熵值：

$$H = -\sum_{k=1}^{N} p_k \log p_k$$

谱熵特征实例

- 第三幅图表示： 1 －熵值



短时过零率特征

过零率定义

- 信号的波形曲线在单位时间内穿过幅度零点的次数

$$Z_n = \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]| \cdot w(n-m)$$

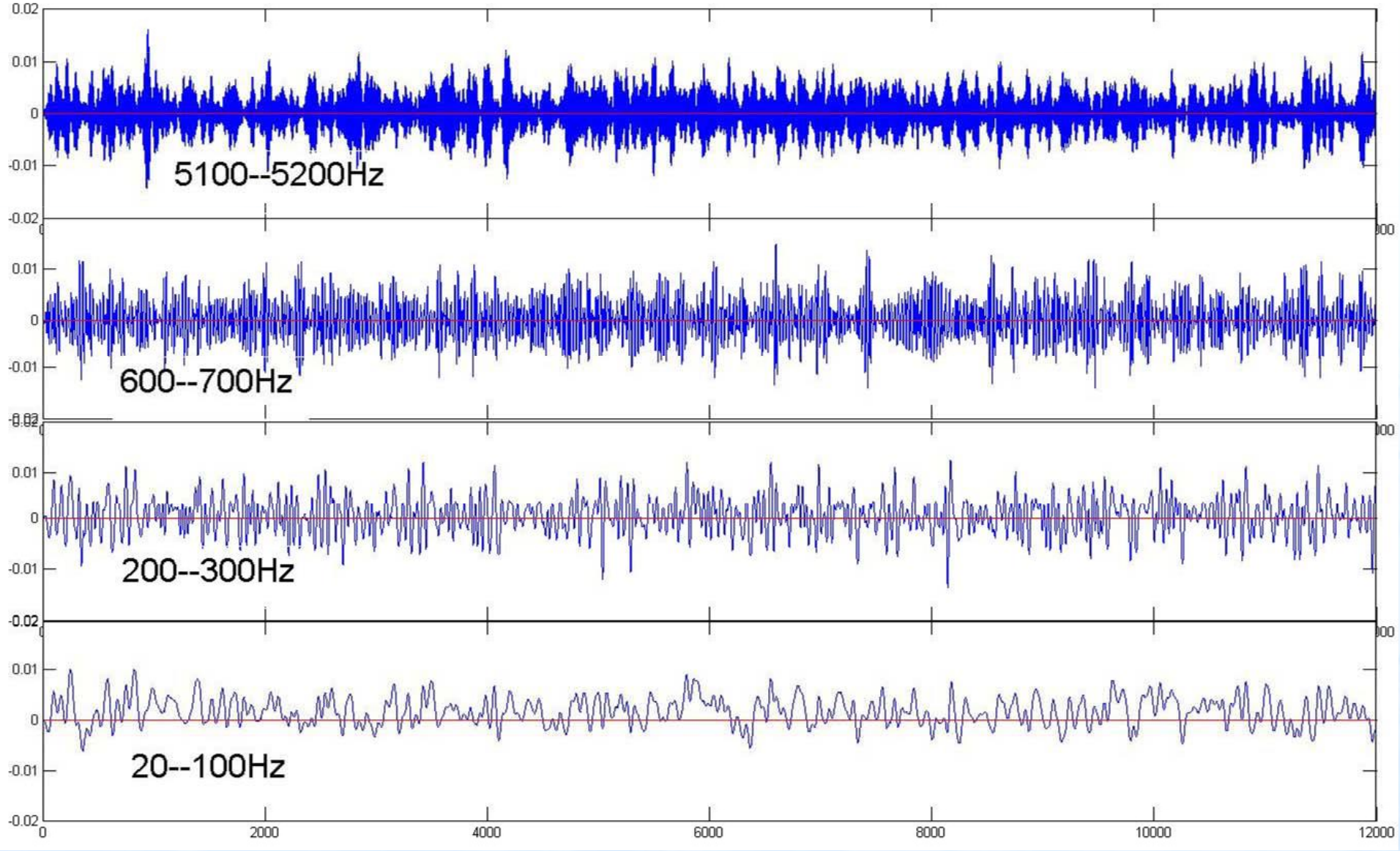
一般取

$$w(n) = \begin{cases} \dfrac{1}{2N} & 0 \le n \le N-1 \\ 0 & \text{其他} \end{cases}$$

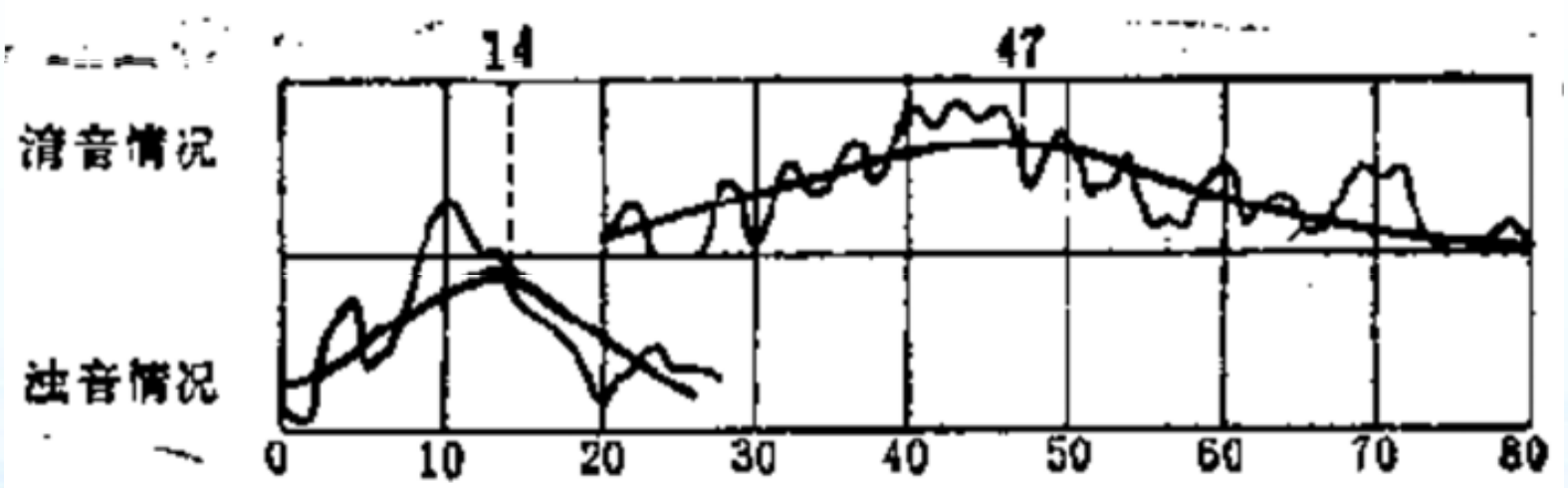
物理意义

- 过零率是区分清音与浊音、有声与无声的重要标志，过零率高的区段对应于清音或无声区（噪声相对较高），过零率低的区段对应于浊音；
- 本质上反映高低频分量的比例。

过零率的物理意义


5100--5200Hz
600--700Hz
200--300Hz
20--100Hz
0.02
0.01
0
-0.01
-0.02
0
2000
4000
6000
8000
10000
12000

过零率对于语音的意义

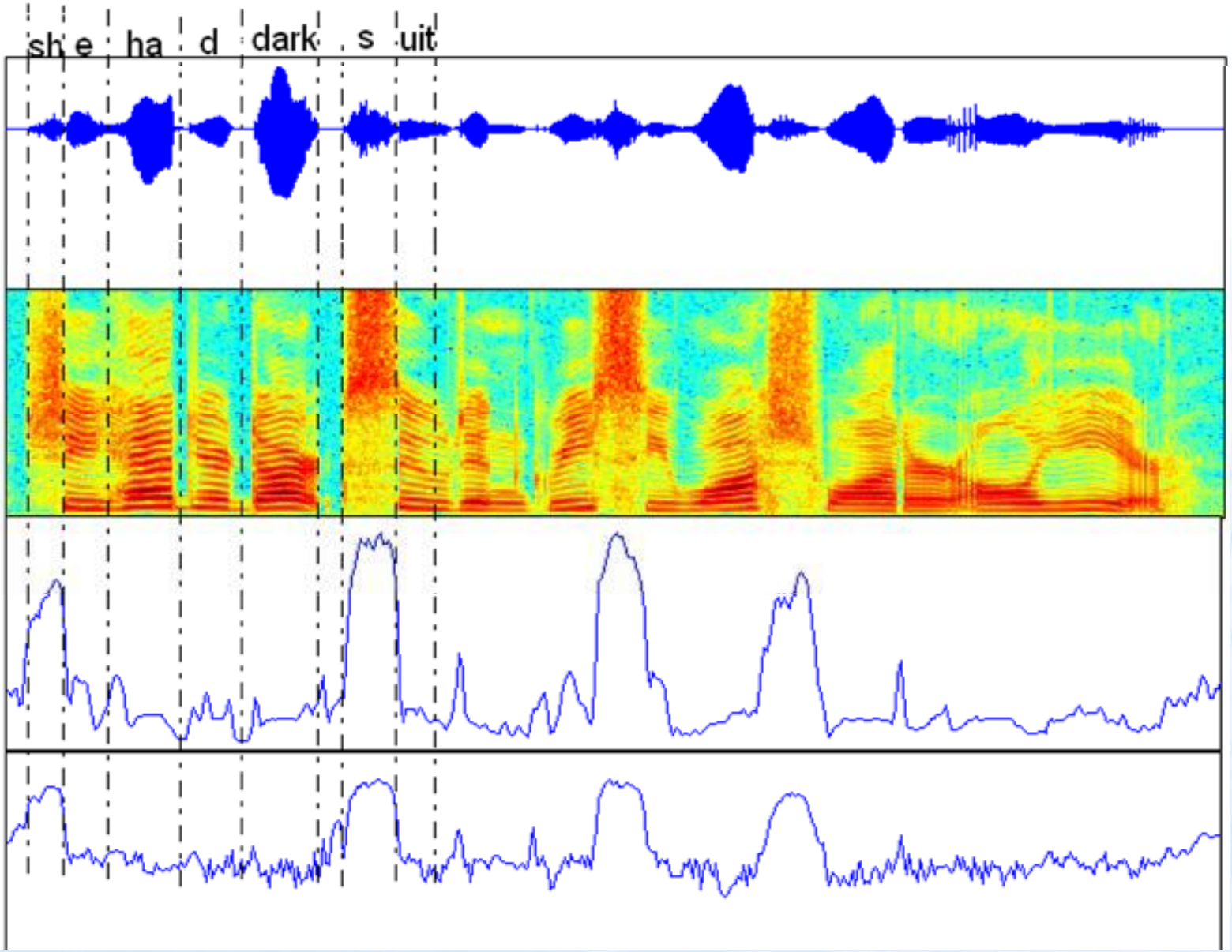


每１０ｍｓ内的过零次数

浊音和清音情况下典型的平均过零率的直方图

- 直方图的分布形状与高斯分布很吻合，而且浊音时的短时平均过零率的均值为１４过零／１０ｍｓ，清音时短时过零率的均值为４７过零／１０ｍｓ。注意到浊音和清音有一个交叠区域，此时很难分清是浊音还是清音，尽管如此，平均过零率仍可以粗略的判断清音和浊音。

过零率实例



时域波形

幅度谱

短时平均过零率

高低频能量比

短时能量特征

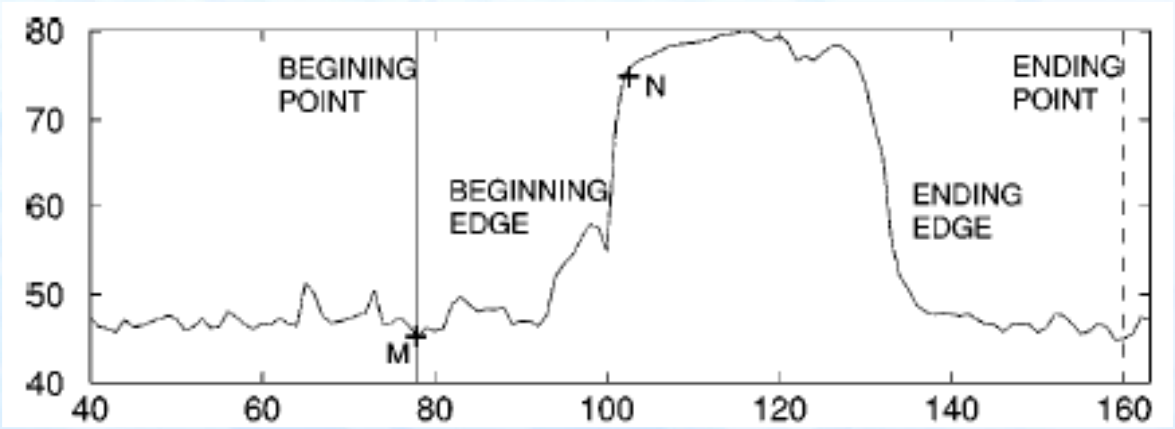
- 能量表达方式

$$magnitude = \sum_{i=1}^{N} |x[i]| \qquad energy = \sum_{i=1}^{N} (|x[i]|)^2$$

- 对数表达方式

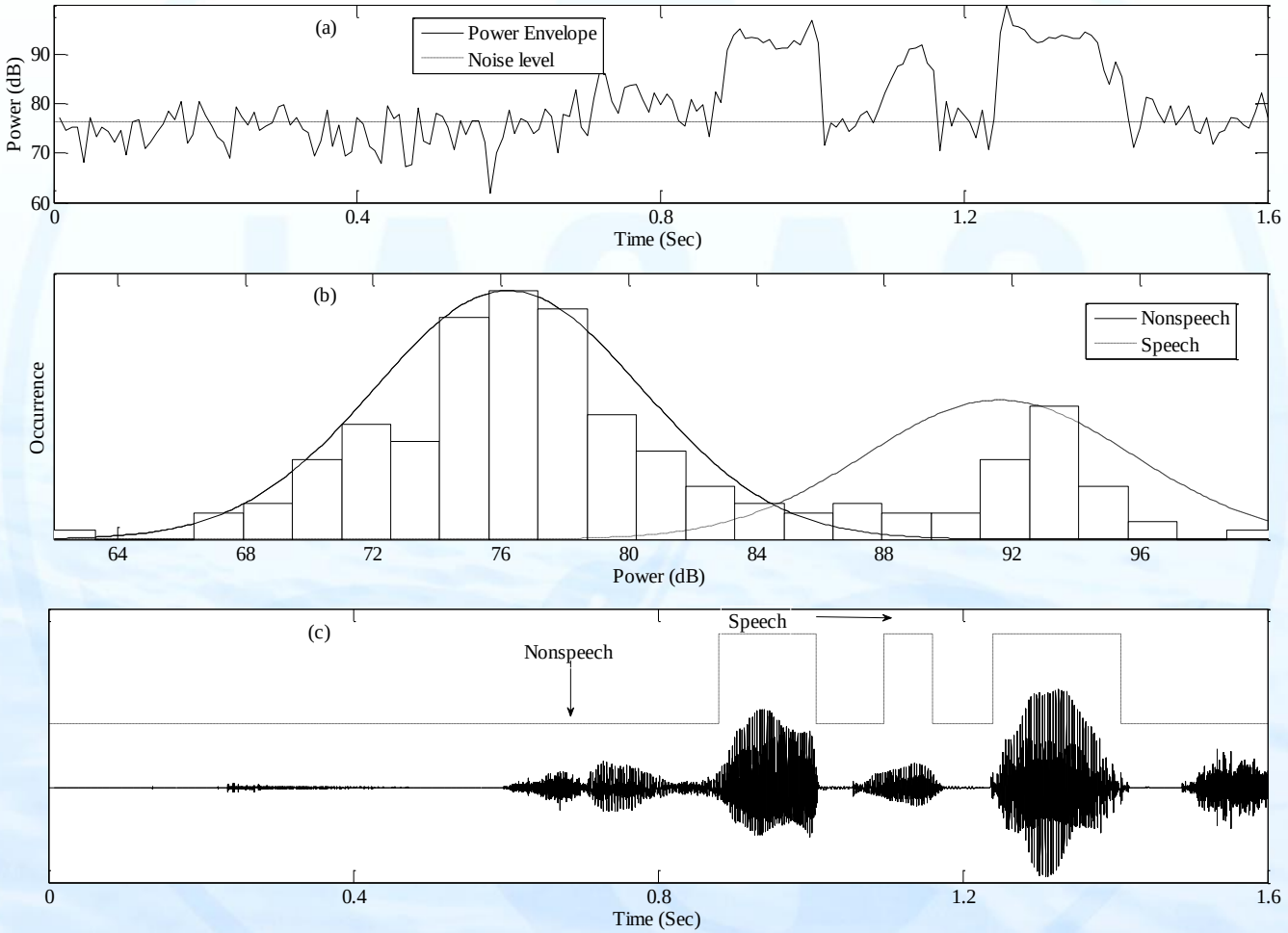
$$db = 10\log_{10} \sum_{i=1}^{N} |x[i]|^2$$

- 能量包络实例



- 能量门限
  - 门限值＝最小值＋ｃｆ＊（平均值－最小值）
  - ｃｆ为０到１之间的小数

短时能量特征


(a)
Power Envelope
Noise level
Power (dB)
90
80
70
60
0
0.4
0.8
1.2
1.6
Time (Sec)
(b)
Nonspeech
Speech
Occurrence
64
68
72
76
80
84
88
92
96
Power (dB)
(c)
Speech
Nonspeech
0
0.4
0.8
1.2
1.6
Time (Sec)

短时能量特征

- 能量用途
  - 可以作为区分清音和浊音的特征参数；
  - 在信噪比比较高的情况下，短时能量可以作为区分有声和无声的依据
  - 可以作为辅助的特征参数用于语音识别中。

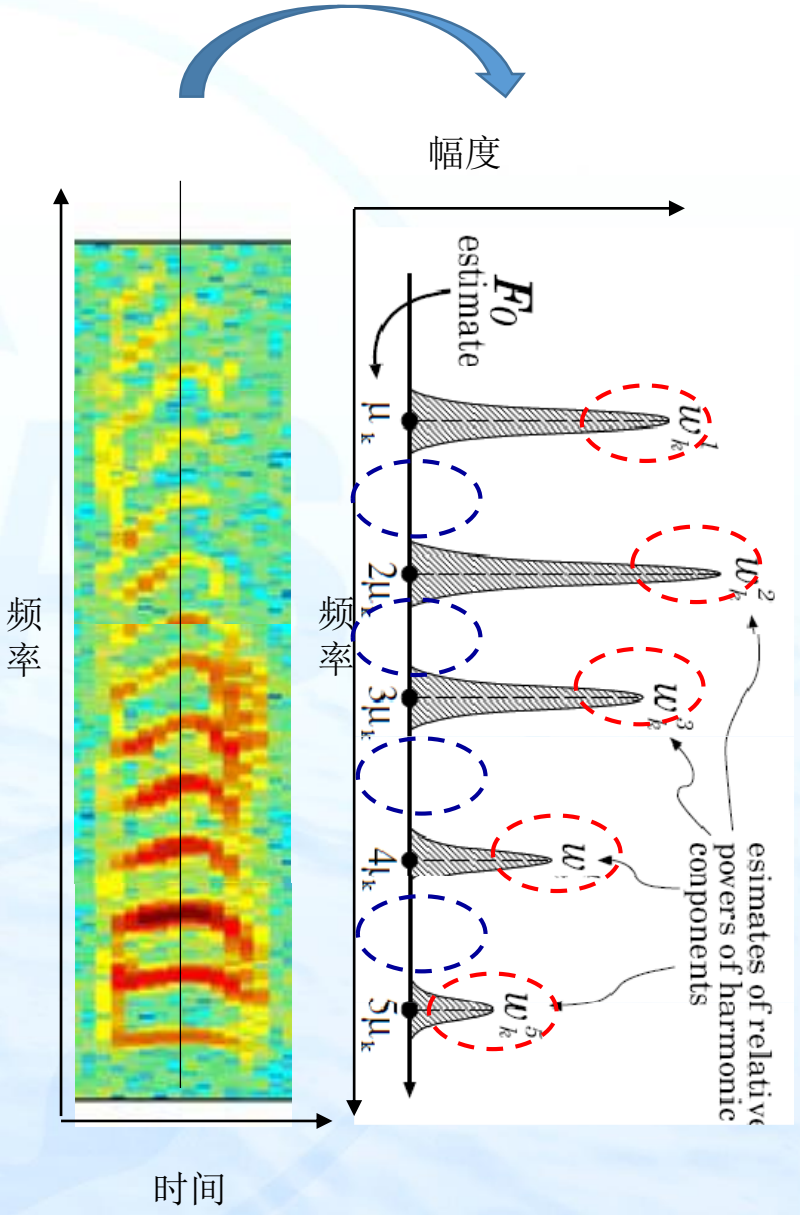
谐波特征

谐波产生

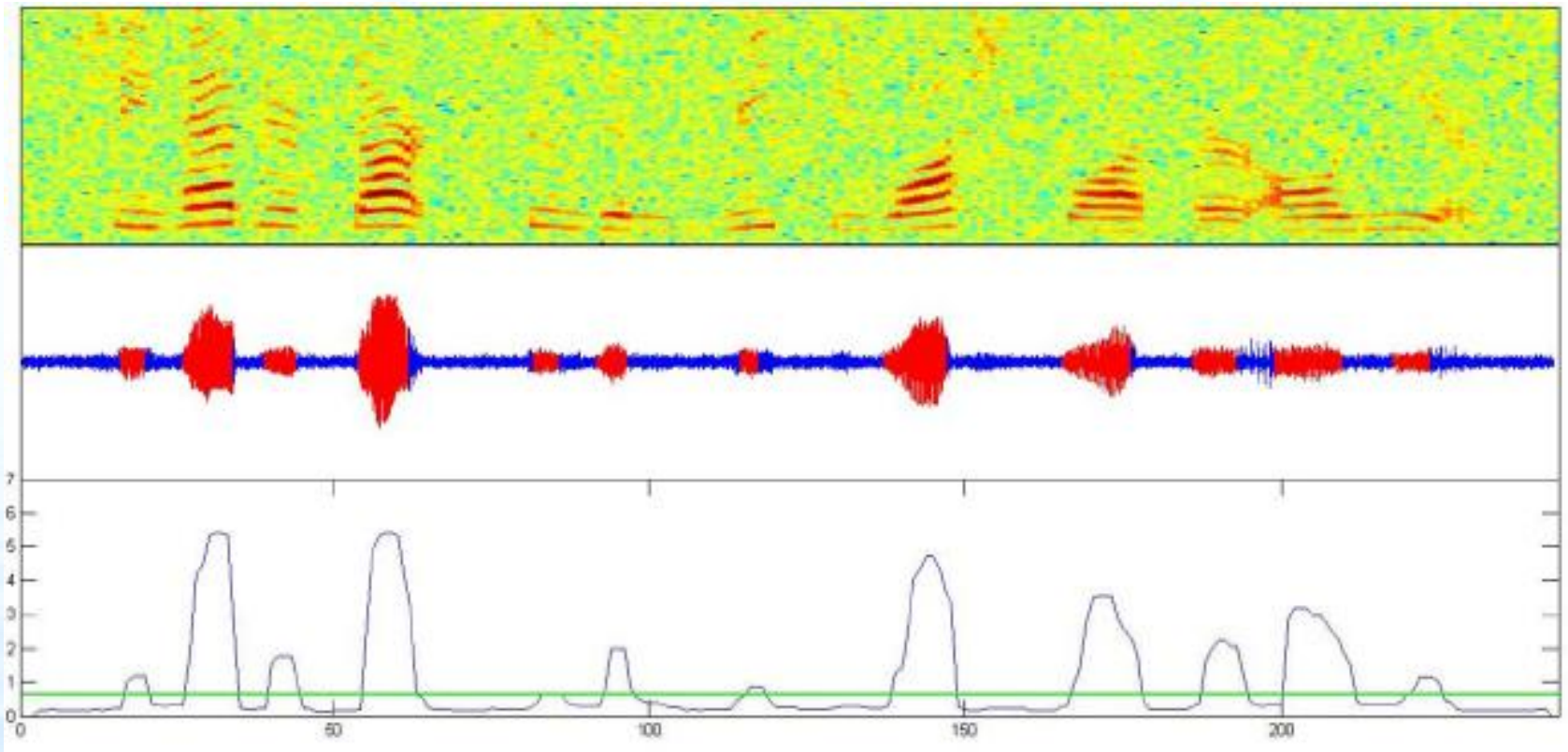
- 气流通过绷紧的声带时，激励声带产生振动，使得声门处形成准周期性的脉冲串，激励声道。声带绷紧的程度决定了震动频率，即基音频率。浊音不仅包含元音，还包含浊辅音。它在时域是准周期的，在频域具有谐波结构。发清音时，声带松弛而不振动，气流通过声门直接进入声道，没有谐波结构。

谐波特征描述

- 确定基频F_0，找到顶峰、谷底。
- 描述值＝顶峰处能量 / 谐波谷底能量

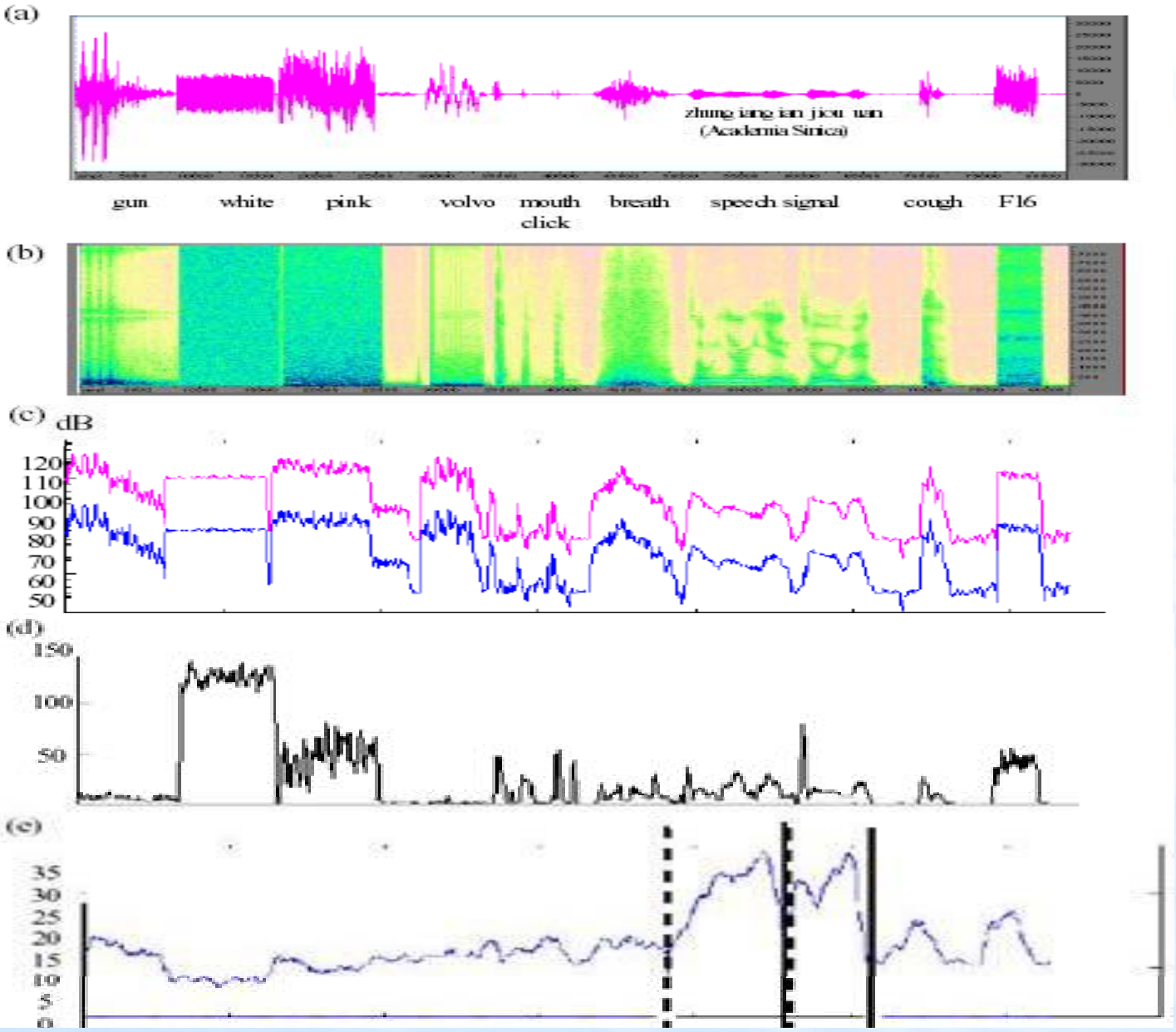


谐波特征实例



1. 谐波结构越显著，特征值越大。
2. 谐波结构能够准确区分浊音与非周期性噪声，对周期性噪声无效。
3. 由于清辅音缺乏谐波结构，谐波特征容易将清辅音丢弃。
4. 谐波结构能够区分语音信号与非稳定的噪声信号。

几种短时特征的实例对比



波形信号

幅度谱

短时能量特征

过零率特征

谱熵特征

总结

- 短时能量
  - 计算方法　$E = \sum x_n^2$
  - 静音和语音、清音和浊音之间均有能量差别
- 短时过零率
  - 定义：一帧语音内波形穿过零电平的次数
  - 清音过零率一般高于浊音
- 短时能量和过零率结合可以导出经典的语音端点检测算法：双门限法

总结

- 语音端点检测（Voice Activity Detection, VAD）
  - 目标：判断语音信号中的语音段落的起始点和终结点的位置
  - 作用：可以用于去掉多余的非有声信号，提高系统处理语音的速度，同时减少因非有声信号进入后端分析系统而产生的干扰。

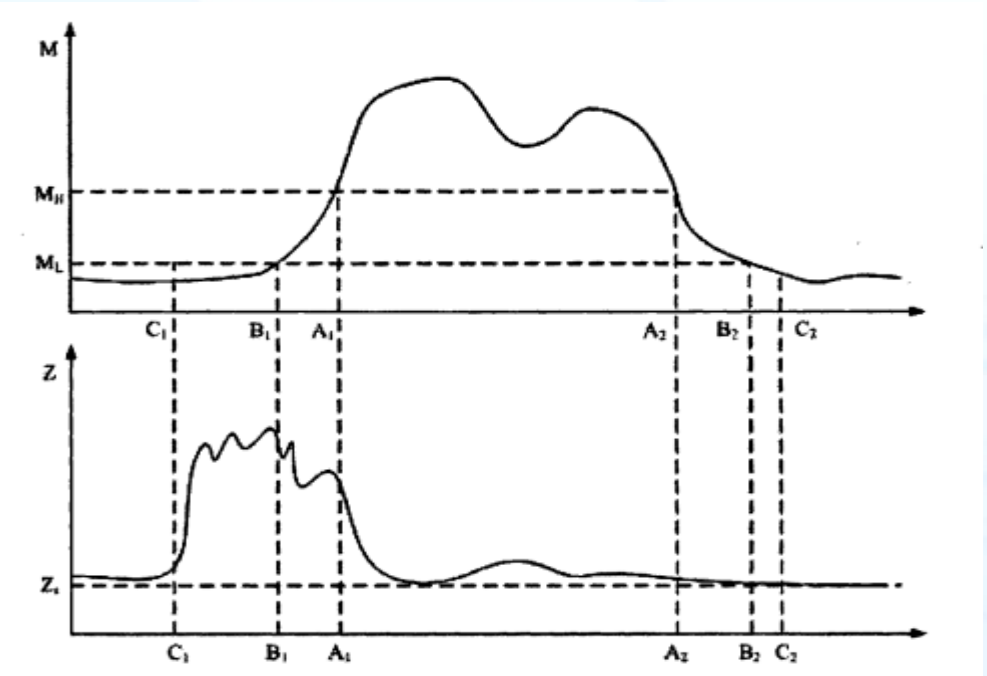
总结

- 方法理论依据
  - 语音信号一般可分为无声段、清音段和浊音段，三者在能量和过零率上具有一定的差异

类型	能量	过零率
无声	低	低
清音	较低	高
浊音	高	低

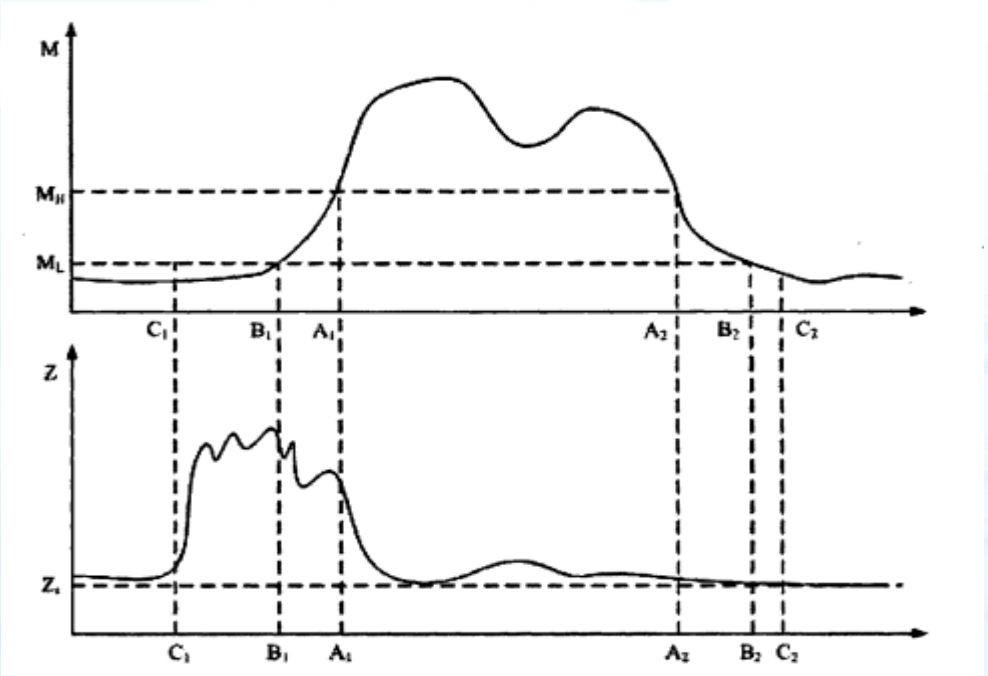
总结

- 步骤1：定位浊音
  - 确定较高的能量限M_H
  - 语音中高于M_H的部分可以基本判定为浊音
  - 确定端点A1和A2



总结

- 步骤2：扩展搜索
  - 选取低能量门限M_L并由A1和A2向两侧扩展，确定B1和B2
  - 确定无声过零率均值Zs，以3倍Zs为门限由B1和B2再次向两侧扩展确定C1和C2，为最终分割结果

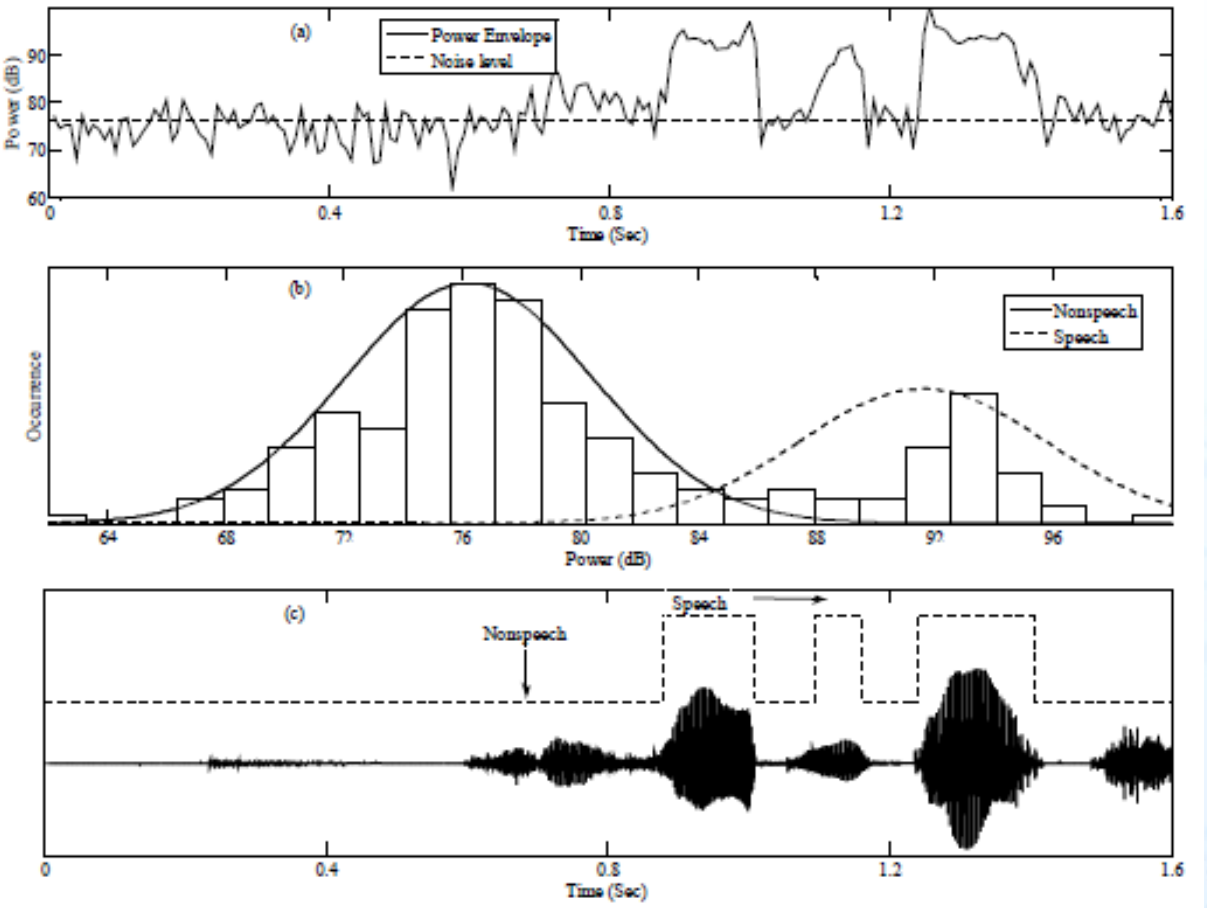


提纲

- 简介
- 原理
- 基于短时声学特征的方法
- 基于能量特征与统计模型的方法
  - 基于时序不相关统计模型的方法
  - 基本时序相关统计模型的方法
- 基于多维特征与机器学习的方法

子带包络的统计特点

- 由于语音是纯净语音和噪声的叠加，所以语音均值大于非语音均值
- 由于通常假设噪声比语音信号更加稳定，所以语音方差大于非语音方差

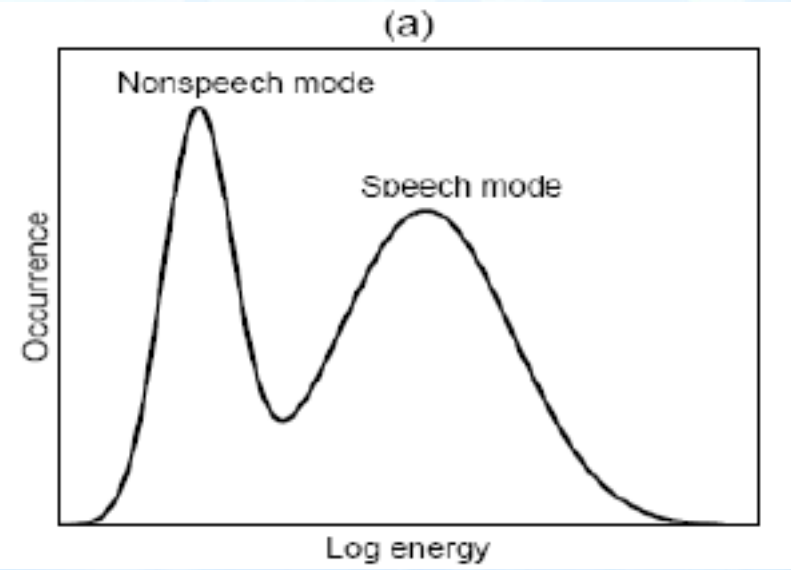


统计模型的表示

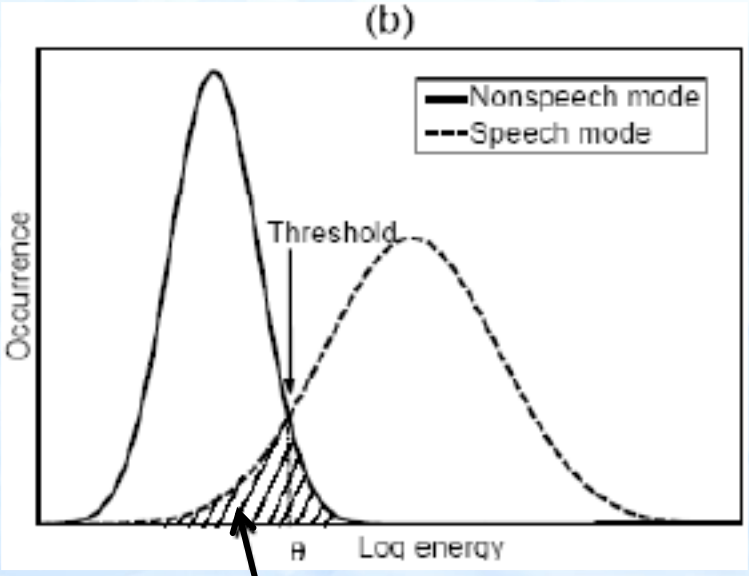
模型的抽象

- 阴影部分表示分类误差
- 门限最小化分类误差
- 模型参数集：均值、方差等等

混合分布＝噪声ＰＤＦ＋语音ＰＤＦ



噪声ＰＤＦ与语音ＰＤＦ



分类误差

面临的问题

- **参数估计**
  - 基于经验规则的估计方法
  - 基于极大似然的最优估计
- **噪声不稳定，统计特征随着时间缓慢变化**
  - 模型参数对观察数据的在线自适应。

基于GMM的建模方法

- 对于某个子带包络上一个观察值x_k，它对应的状态 z 用 0 和 1 表示，$z \in \{0,1\}$， 1 表示语音， 0 表示非语音。λ 表示模型对应的参数集。
- 根据贝叶斯原理，它的概率密度函数可以表示为：

$$p(x_k \mid \lambda) = \sum_z p(x_k, z \mid \lambda) = \sum_z p(x_k \mid z, \lambda) p(z)$$

其中$p(z=0)$表示非语音出现的先验概率，$p(z=1)$表示语音出现的先验概率。 $p(z=0)+p(z=1)=1$

$p(x_k \mid z=0, \lambda)$ 表示给定参数集的情况下，样本观察值为非语音的似然度。

$p(x_k \mid z=1, \lambda)$ 表示给定参数集的情况下，样本观察值为语音的似然度。

$$p(x_k \mid z, \lambda) = \frac{1}{\sqrt{2\pi\kappa_z}} \exp\{-(x_k - \mu_z)^2 / 2\kappa_z\}$$

基于GMM的建模方法

- 对于某个子带包络 $x=\{x_1, x_2, ..., x_M\}$，它出现得似然度，也就是概率密度函数表示为：

$$p(x \mid \lambda)=\prod_{k=0}^{M} p(x_k \mid \lambda)$$

- 参数的极大似然估计表示为：

$$\hat{\lambda}=\max_{\lambda} \ln p(x \mid \lambda)$$

- 根据求得的最优参数集 λ，可以进一步求解分类门限。

$$p(\theta \mid z=1, \lambda) p(z=1)=p(\theta \mid z=0, \lambda) p(z=0)$$

可以证明：分类门限使得分类误差最小。

基于ＥＭ算法的参数估计

- 从子带包络

  $x=\{x_1,x_2,.....,x_M\}$

  推导模型的最优参数估计。

- ＥＭ算法的基本原理：假定已知一个参数集λ'，求解一个更优的参数集λ。算法迭代进行，直至收敛。

- 数学上可以证明，每次迭代都将导致似然度$p(x|\lambda)$不断上涨，直到最大值。

- 这里 $w_0=p(x_k|z=0,\lambda)$

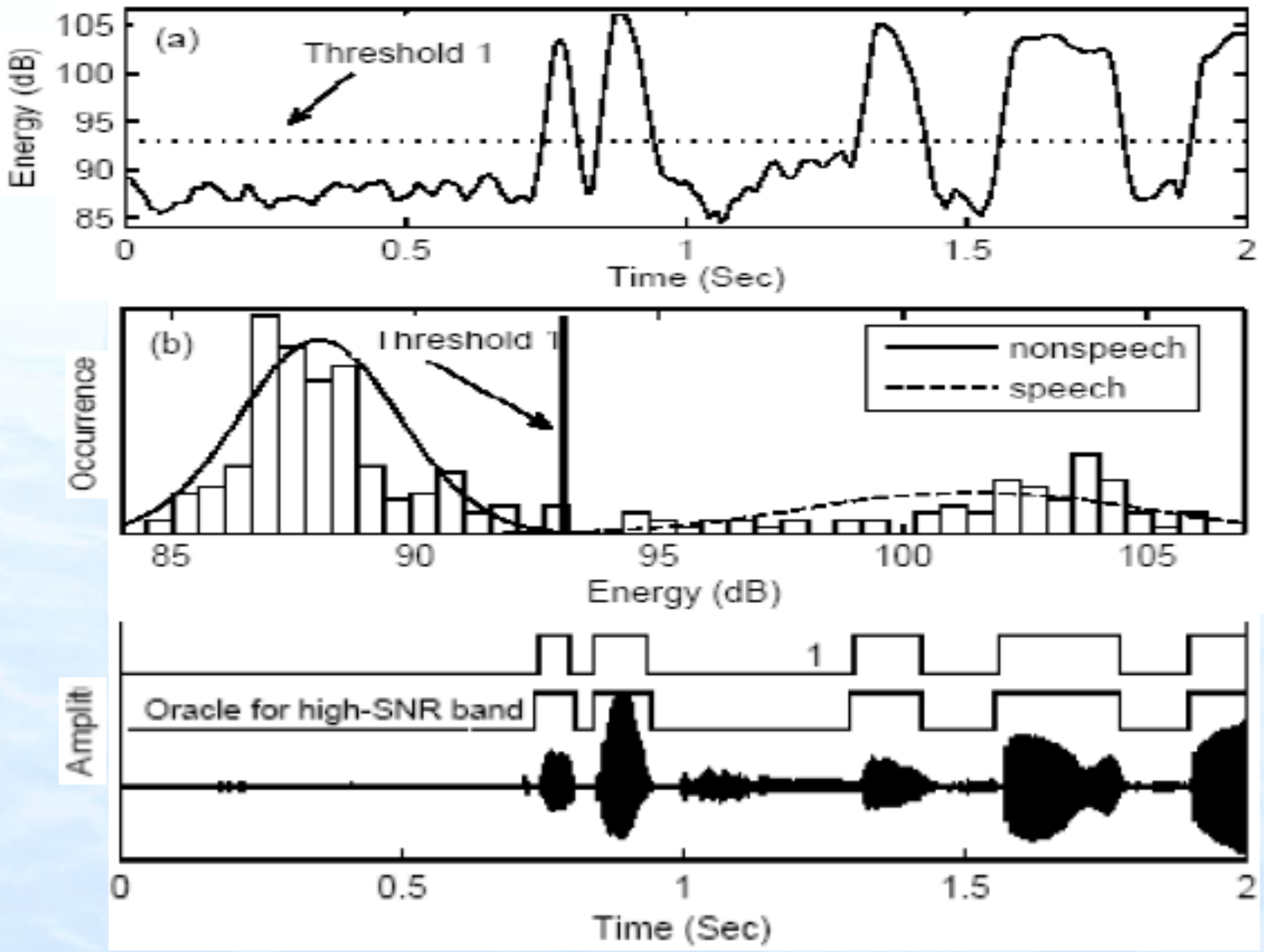
$$\bar{w}_z=\frac{1}{M+1}\sum_{k=0}^{M}p(z|x_k,\lambda')$$

$$\bar{\mu}_z=\frac{\sum_{k=0}^{M}x_k\,p(z|x_k,\lambda')}{(M+1)\bar{w}_z}$$

$$\bar{\kappa}_z=\frac{\sum_{k=0}^{M}(x_k-\bar{\mu}_z)^2\,p(z|x_k,\lambda')}{(M+1)\bar{w}_z}$$

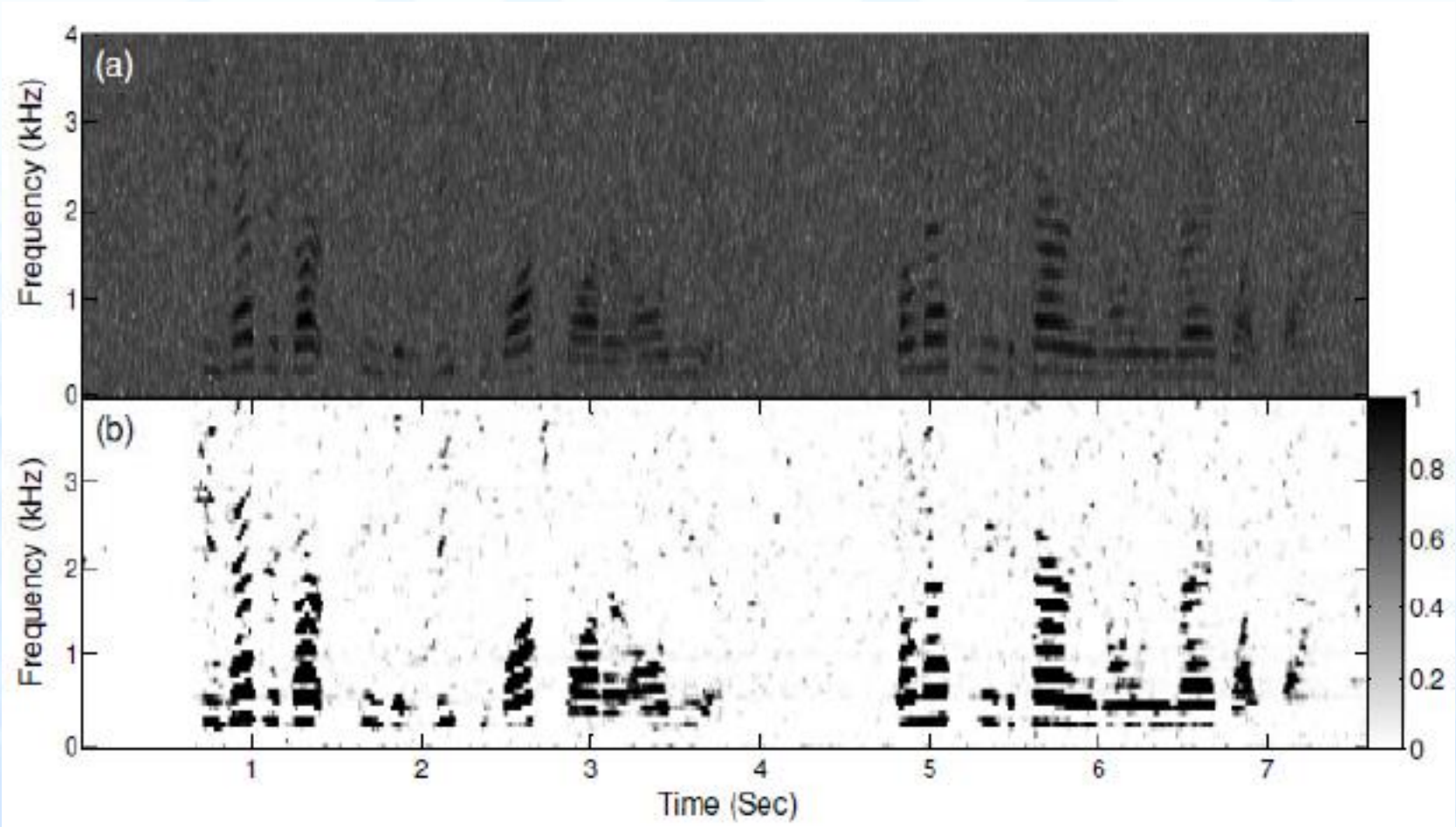
$$p(z|x_k,\lambda')=\frac{w'_z\,p(x_k|z,\lambda')}{\sum_z w'_z\,p(x_k|z,\lambda')}$$

实例演示


(a)
Threshold 1
Energy (dB)
105
100
95
90
85
0
0.5
1
1.5
2
Time (Sec)
(b)
Threshold 1
nonspeech
speech
Occurrence
85
90
95
100
105
Energy (dB)
1
Oracle for high-SNR band
Amplit
0
0.5
1
1.5
2
Time (Sec)

时频域语音出现概率

$$p(z=1|\,x_k,\lambda)=\frac{p(z=1)\,p(x_k\,|\,z=1,\lambda)}{p(z=1)\,p(x_k\,|\,z=1,\lambda)+p(z=0)\,p(x_k\,|\,z=0,\lambda)}$$



算法的在线化

- EM算法是个离线过程，需要一整段语音数据估计参数。在最终结果出现前需要等待一段时间，产生了时间延迟。
- 在一些实时应用（例如语音通信）中，无法忍受时间延迟，需要根据当前的观察值立即做出决策。
- 均值、方差、先验概率需要表达为当前的观察值，与上一时刻参数集的函数。

$$\lambda_{k+1} = f(\lambda_k, x_{k+1})$$

EM算法的在线化处理

- 均值

$$\mu_{k+1,z} = \frac{\alpha w_{k,z}\mu_{k,z} + (1-\alpha)p(z \mid x_{k+1}, \lambda_k)x_{k+1}}{w_{k+1,z}}$$

- 方差

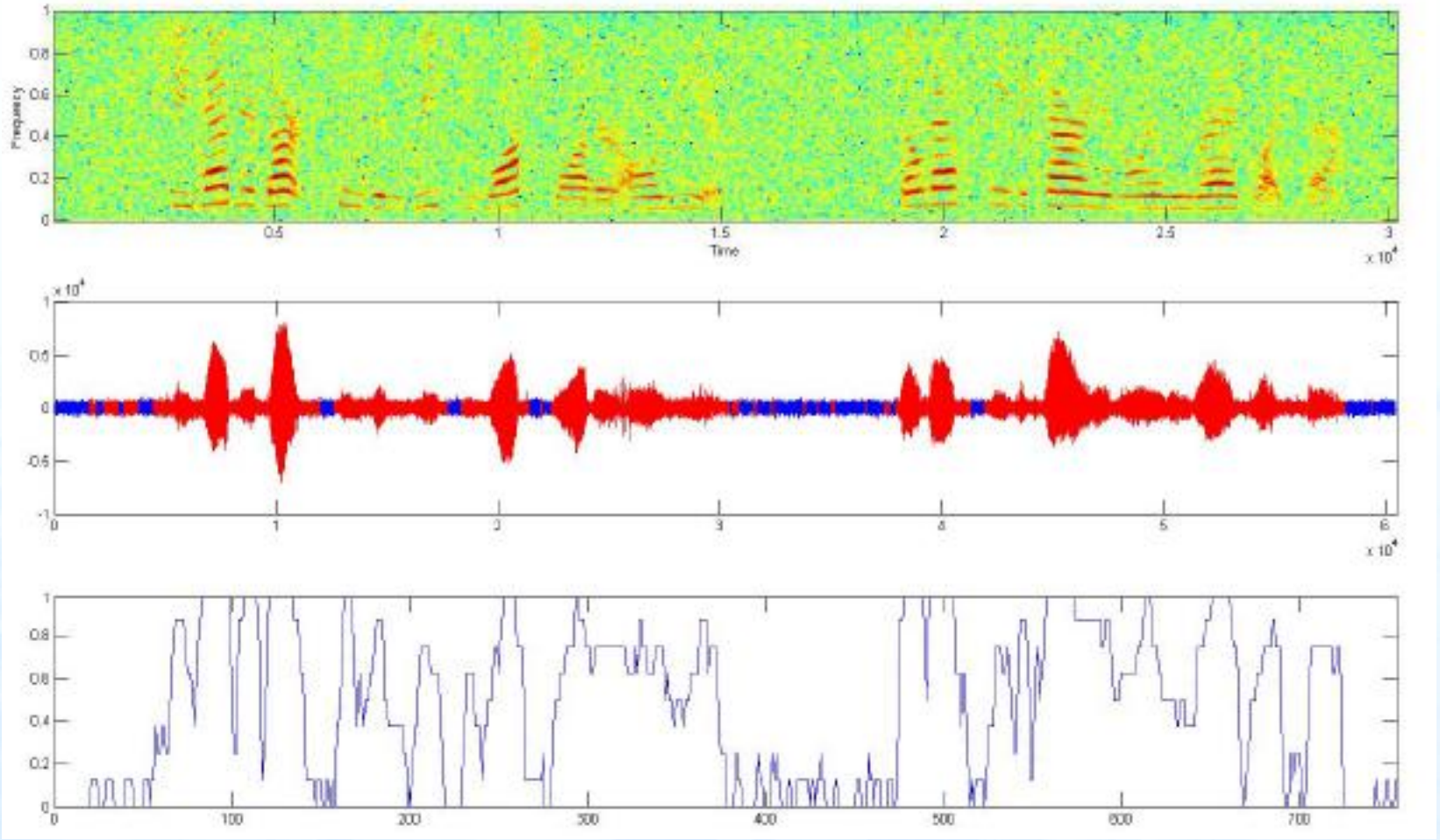
$$\kappa_{k+1,z} = \frac{\alpha w_{k,z}\kappa_{k,z} + (1-\alpha)p(z \mid x_{k+1}, \lambda_k)(x_{k+1} - \mu_{k+1,z})^2}{w_{k+1,z}}$$

- 先验概率

$$w_{k+1,z} = \alpha w_{k,z} + (1-\alpha)p(z \mid x_{k+1}, \lambda_k)$$

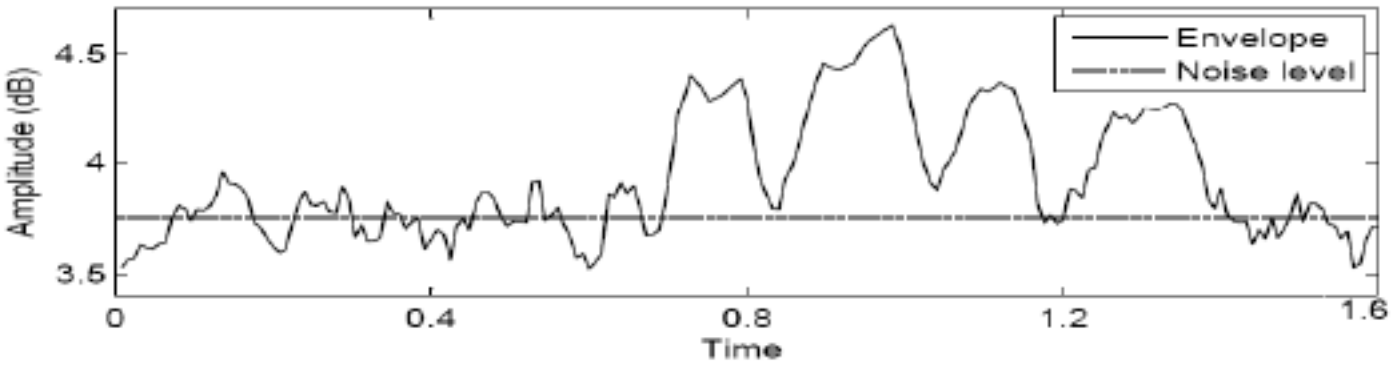
这里的 $\alpha = 0.98$ 表示忘记因子，k 表示时间索引。

决策过程实例演示

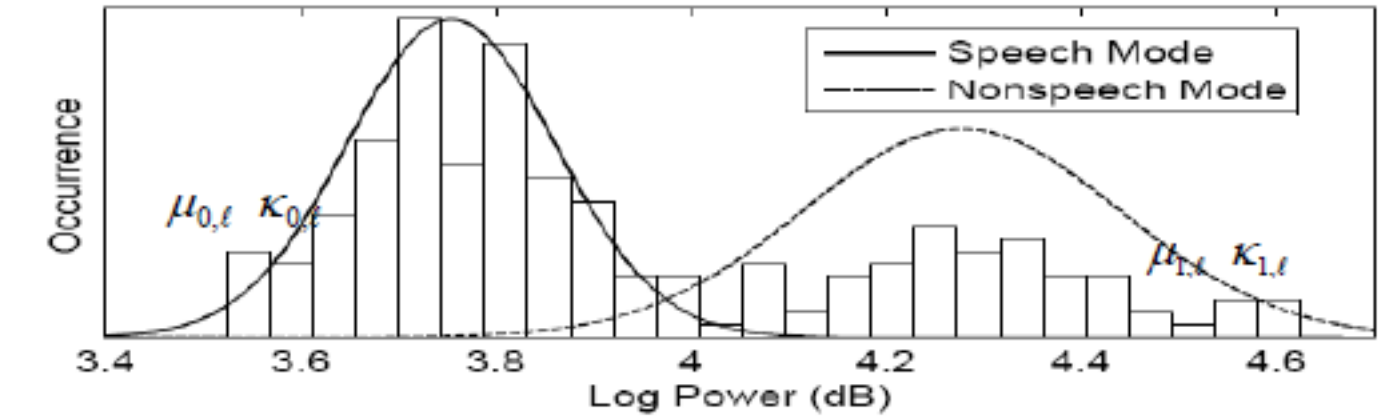


语音 / 非语音段建模

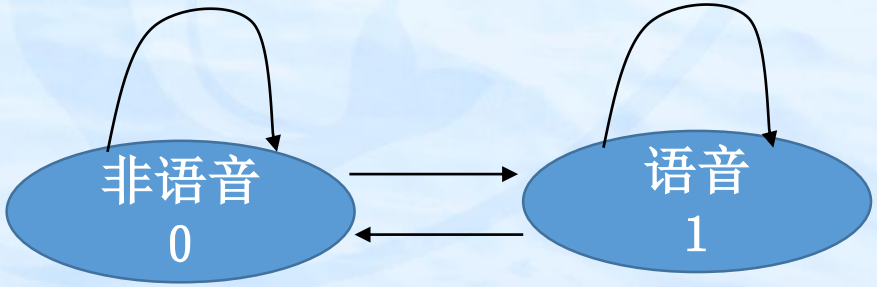
功率谱包络序列



功率谱直方图



二元马尔科夫链



二元马尔科夫模型的在线化

- 在线化的目的
  - 在线估计，使输出结果延迟为零
- 在线似然函数的定义

$$Q_{\ell+1|\lambda_\ell}(\lambda) \triangleq E\{\log p(\mathrm{x}_{\ell+1}, \mathrm{s}_{\ell+1} \mid \lambda) \mid \mathrm{x}_{\ell+1}, \lambda_\ell\}$$

ℓ：时间索引

观察值序列：

$$\mathrm{x}_\ell = \{x_1, x_2, \ldots, x_\ell\}$$

状态序列：

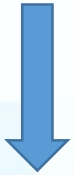
$$\mathrm{s}_\ell = \{s_1, s_2, \ldots, s_\ell\}, \quad s_t \in \{0,1\}$$

参数集：

$$\lambda_\ell = \{\mu_{0,\ell}, \kappa_{0,\ell}, \mu_{1,\ell}, \kappa_{1,\ell}, \alpha_\ell\}$$

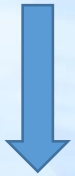
在线估计

$$\lambda_{\ell+1} = \max_{\lambda} \arg Q_{\ell+1|\lambda_\ell}(\lambda) \longrightarrow$$ 似然值最大化



求取一阶导数

$$\begin{cases} \lambda_{\ell+1} = \lambda_\ell + (I_{\ell+1}(\lambda_\ell))^{-1} S_{\ell+1}(\lambda_\ell), \\ I_{\ell+1}(\lambda_\ell) = -\partial^2 Q_{\ell+1|\lambda_\ell}(\lambda) / \partial\lambda^2 \,|_{\lambda=\lambda_\ell} \\ S_{\ell+1}(\lambda_\ell) = -\partial Q_{\ell+1|\lambda_\ell}(\lambda) / \partial\lambda \,|_{\lambda=\lambda_\ell} \end{cases}$$

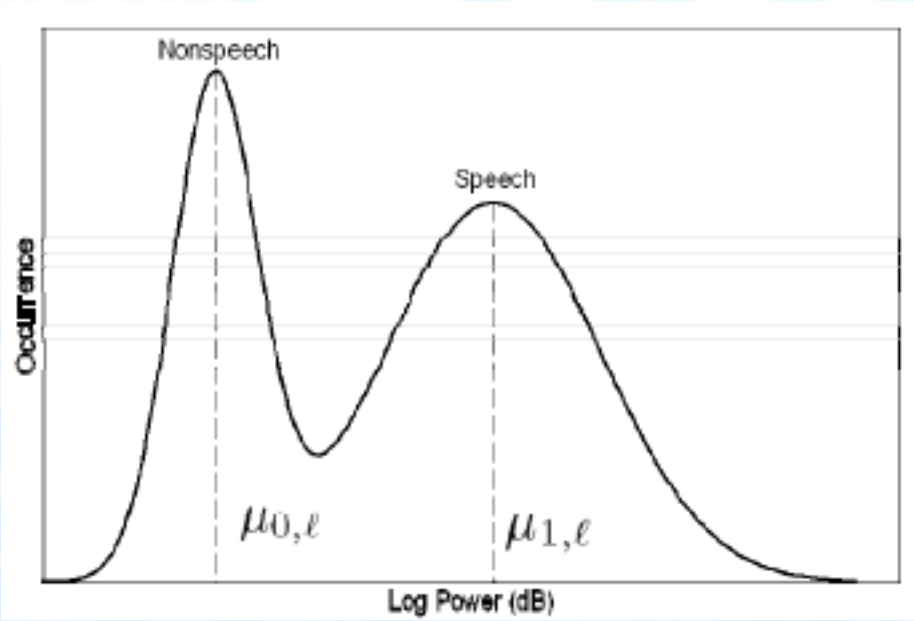


简化为

$$\lambda_{\ell+1} = \varphi(\lambda_\ell, x_{\ell+1})$$

对马尔科夫模型的约束

- 两个分布特征
  - 噪声信号相对于语音信号稳定
  - 语音功率谱在平均意义上大于非语音功率谱



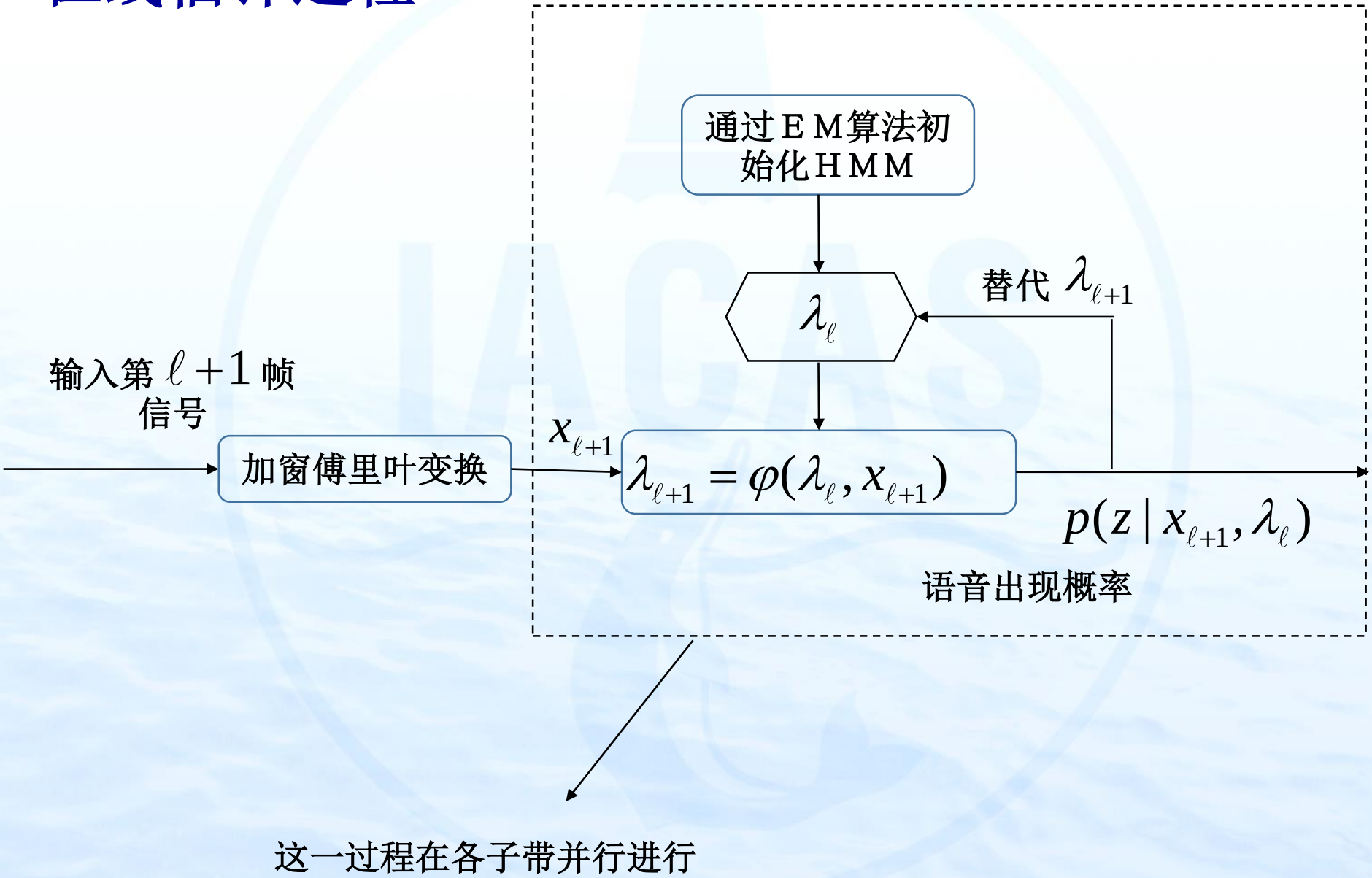
对于HMM的约束

$$\begin{cases} k_{0,\ell} < k_{1,\ell} \\ \mu_{1,\ell} > \mu_{0,\ell} \end{cases}$$

$$\lambda_{\ell+1} = \max_{\lambda} \arg Q_{\ell+1|\lambda_\ell}(\lambda) + \begin{cases} k_{0,\ell} < k_{1,\ell} \\ \mu_{1,\ell} > \mu_{0,\ell} \end{cases}$$

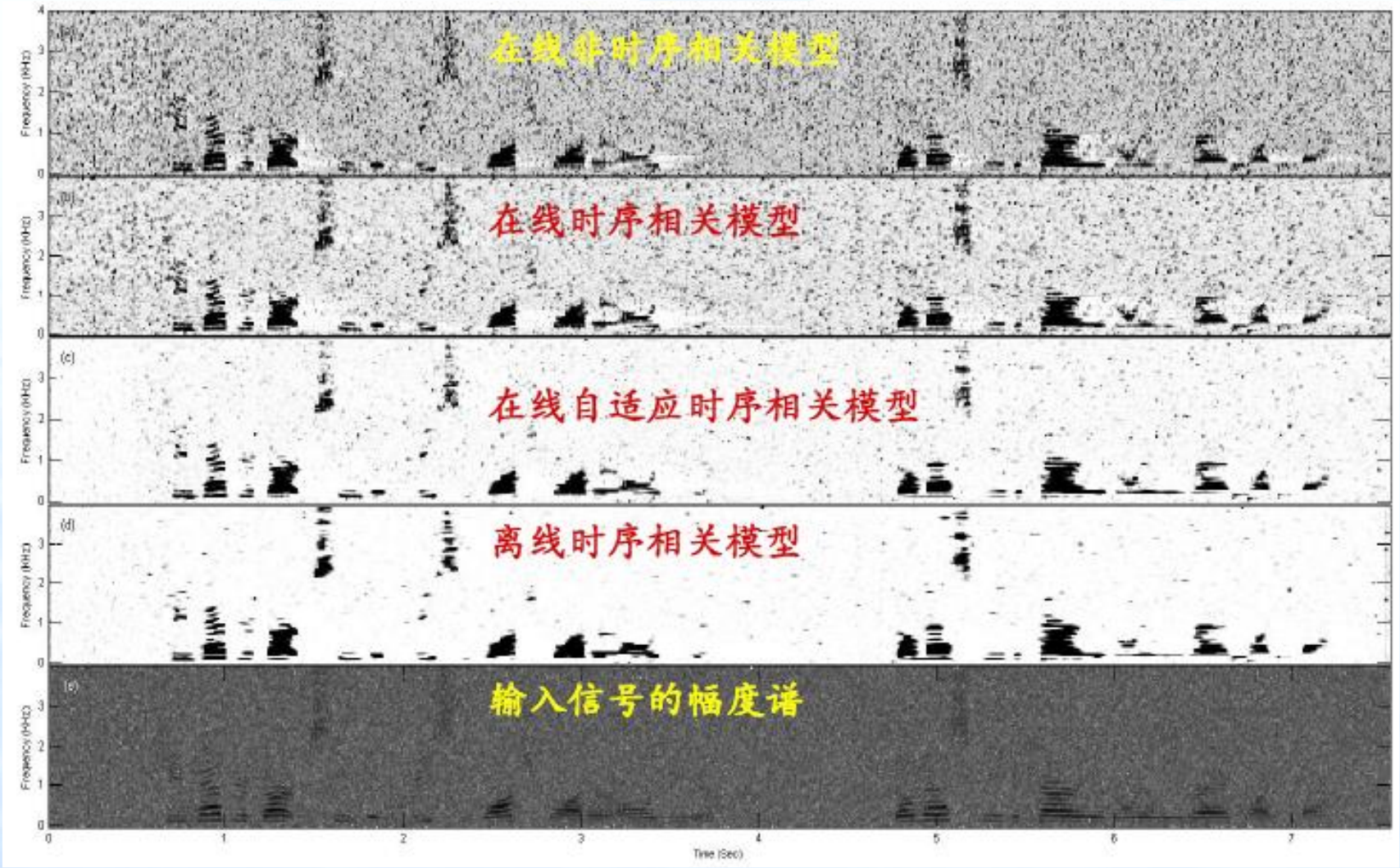
= 约束最大化

在线估计过程


通过ＥＭ算法初始化ＨＭＭ
λ_ℓ
替代 $\lambda_{\ell+1}$
输入第 $\ell+1$ 帧信号
加窗傅里叶变换
$x_{\ell+1}$
$\lambda_{\ell+1}=\varphi(\lambda_\ell, x_{\ell+1})$
$p(z \mid x_{\ell+1}, \lambda_\ell)$
语音出现概率
这一过程在各子带并行进行

时序相关性估计与非时序相关性的比较

- 语音出现概率



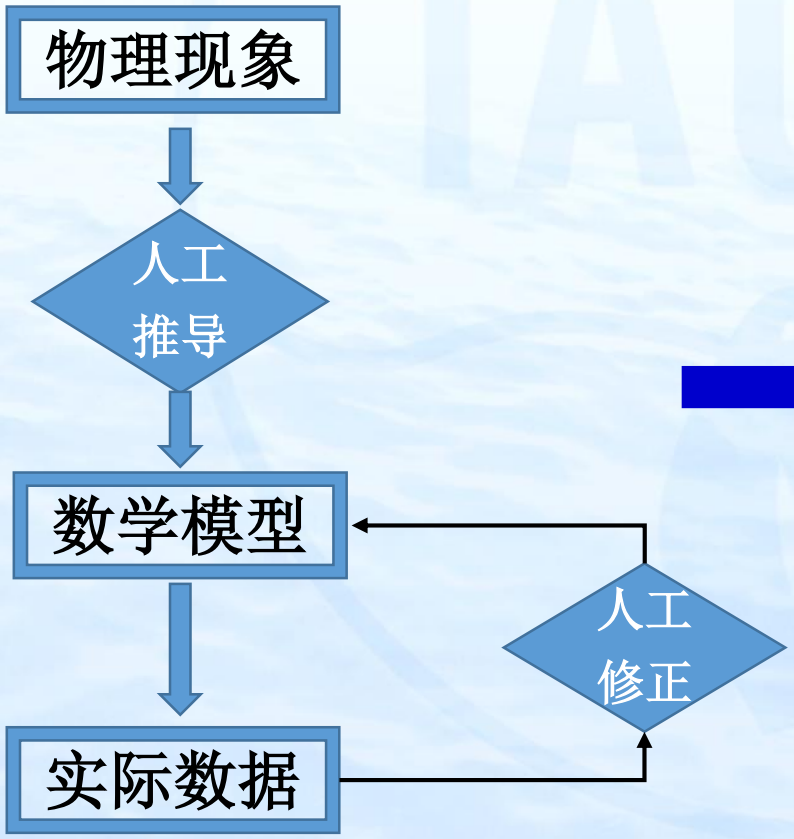
提纲

- 简介
- 原理
- 基于短时声学特征的方法
- 基于能量特征与统计模型的方法
- 基于多维特征与机器学习的方法

深度神经网络的背景

- 深度学习与大数据深刻地改变传统研究领域
  - 传统方法：采集数据，抽取声学特征，训练统计模型。
  - 深度学习：采集数据，特征自学习，训练统计模型。

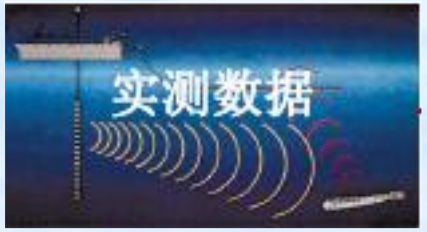
传统方法



大数据
＋
深度学习

机器学习

直接应用



深度神经网络的进展

- 深度学习与大数据的成功案例

语音

相对错误率

３０％

图像

９６％

７４％

准确率

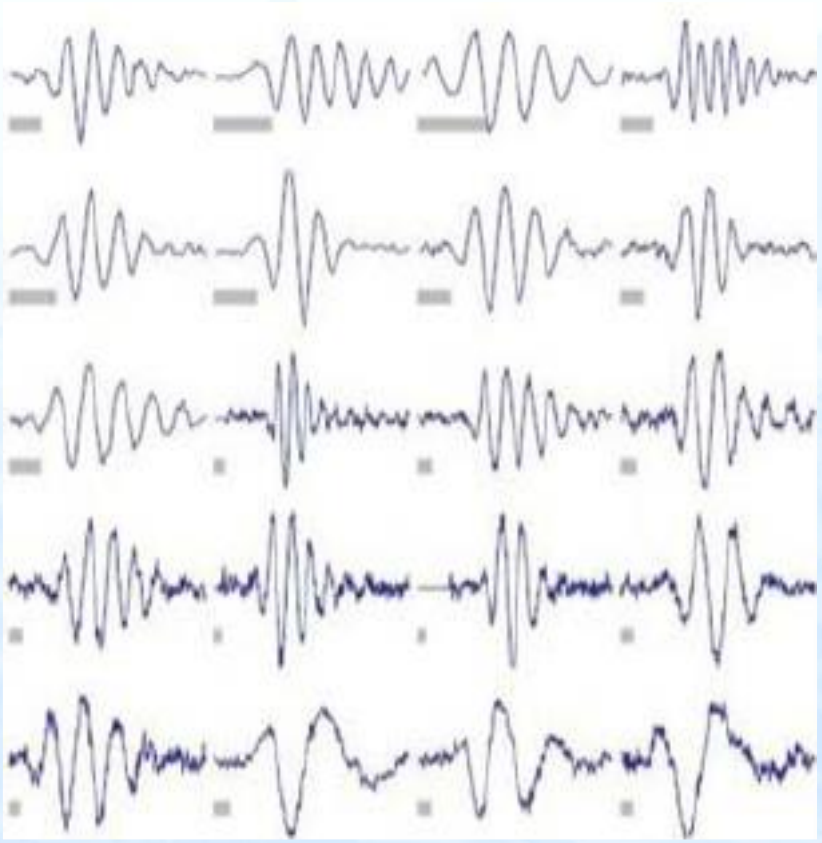
围棋

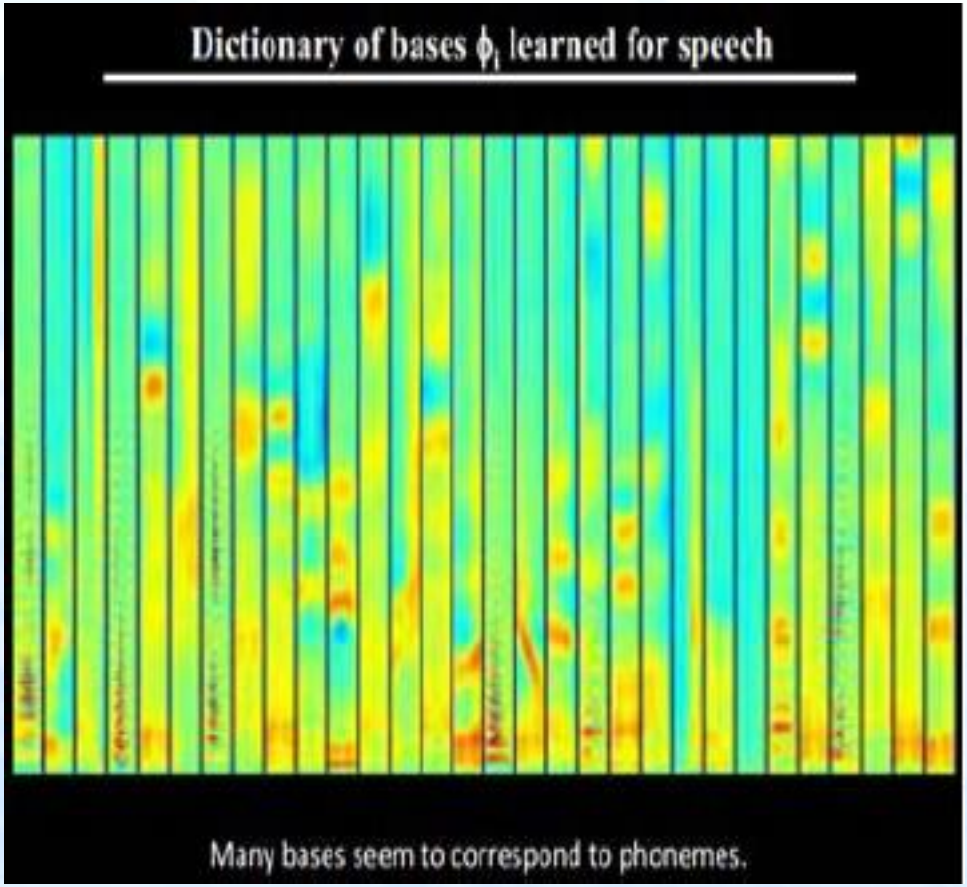
Ａｌｐｈａｇｏ击败李世石

深度学习与大数据推动各个领域的进步幅度超过了许多年进步的总和！

DNN对于语音信号的物理意义

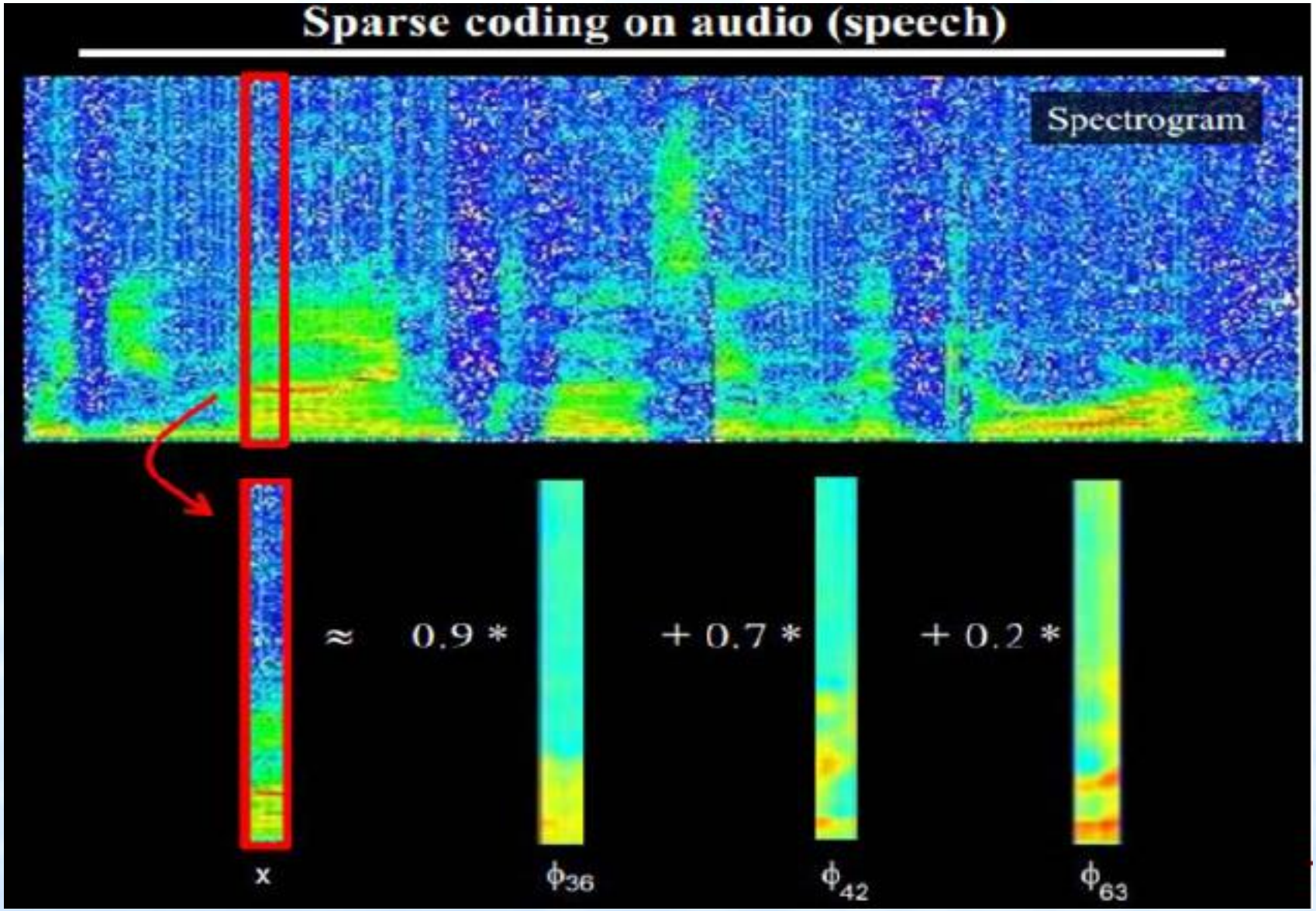
- 语音信号可由20多种基本的语音单元构成，这些基本单元可以看做语音空间的基向量，这20多个基向量展成了语音信号空间。





DNN对于语音信号的物理意义

- 任何语音信号可以由这20多个基向量合成。

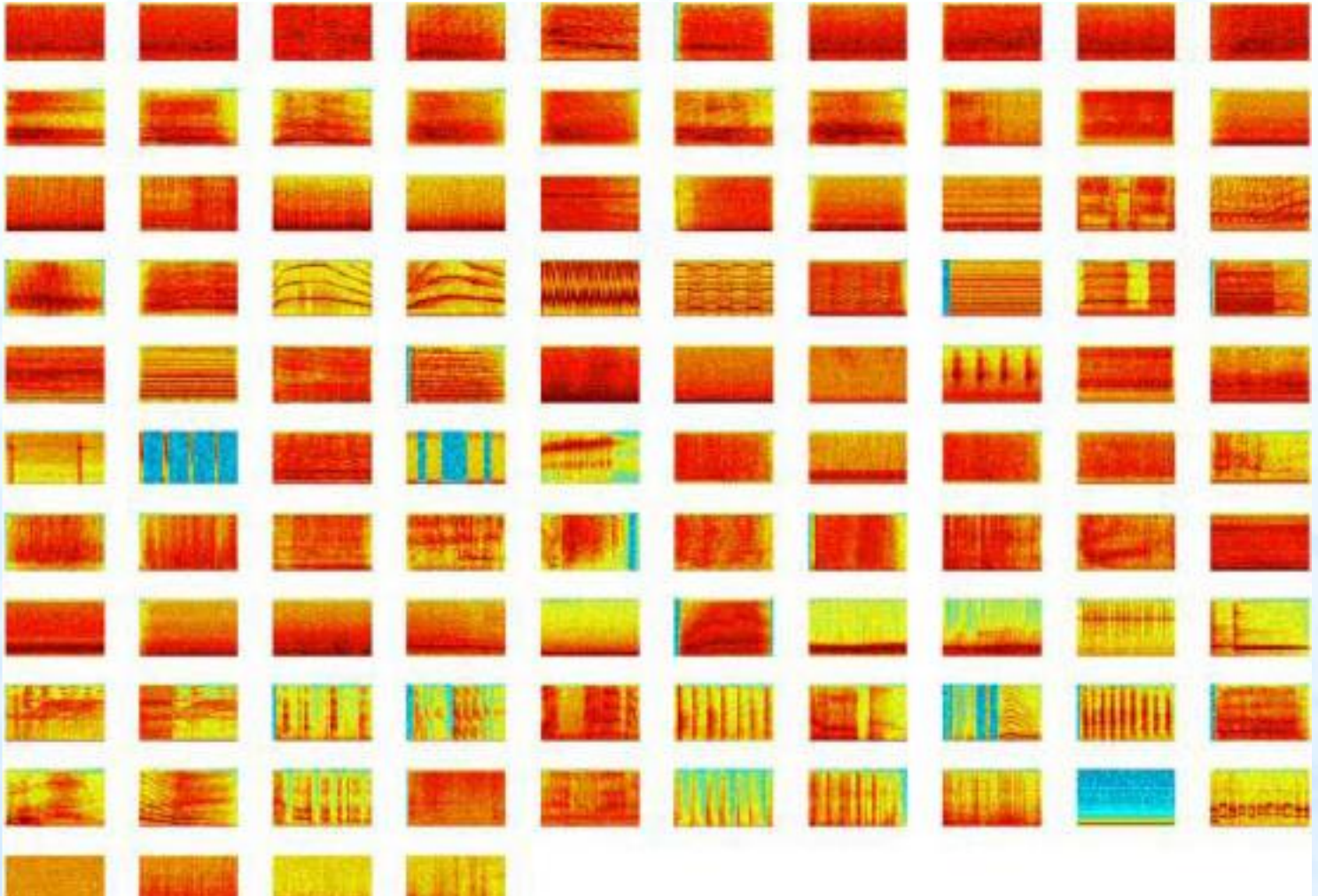


深度学习在ＶＡＤ上的应用

- 语音分帧与快速傅里叶变换；
- 抽取幅度谱，以连续若干帧的幅度谱形成长向量作为声学特征；
- 手工标注语音信号；每个句子中每个时间点和手工标注形成一个训练样本；
- 将纯净语音数据库，与噪声混合，形成大数据；
- 训练神经网络。

DNN的扩展性

- 多条件训练
  - 将纯净语音与各种噪声、各种信噪比混合，对噪声条件进行穷举；



VAD的试验比较

- AUC准确率

Noise	SNR	Sohn	Ramirez05	Ying	SVM	Zhang13	bDNN
Babble	−5 dB	70.69	75.90	64.63	81.05	82.84	**89.05**
	0 dB	77.67	83.05	70.72	86.06	88.33	**91.70**
	5 dB	84.53	87.85	78.70	90.49	91.61	**93.60**
Factory	−5 dB	58.17	58.37	62.56	78.63	81.81	**87.42**
	0 dB	64.56	67.21	68.79	86.05	88.39	**91.67**
	5 dB	72.92	76.82	75.83	89.10	91.72	**93.37**
Volvo	−5 dB	84.43	89.63	92.51	93.91	94.58	**94.71**
	0 dB	88.25	90.44	93.42	93.43	94.80	**95.04**
	5 dB	90.89	90.99	94.13	94.12	95.02	**95.19**

DNN的优缺点

- 优点
  - 辨识语音与非稳定噪声
  - 良好的扩展性
- 缺点
  - 信道的敏感性
  - 时间延迟
  - 计算量大，要求一定的存储空间

谢 谢！

深度学习的VAD算法

优越性：

- 相比于基于支持向量机（SVM）的 VAD 模型深度学习网络具有更强的非线性变换能力，对语音和非语音的分类问题有更强的表达能力 ；
- 另一方面，深度学习神经网络具有特征再学习的能力，可以充分挖掘数据中潜在的信息，避免了专门设计 VAD 特征的困难 。

DNN-VAD

特点：

- 多层全连接的MLP的叠加
- 对最后一层通过softmax函数计算语音/非语音的后验概率。

$$P(C_k|x) = exp(h_{ik}) / \sum_K exp(h_{ik})$$

BDNN-VAD

特点：

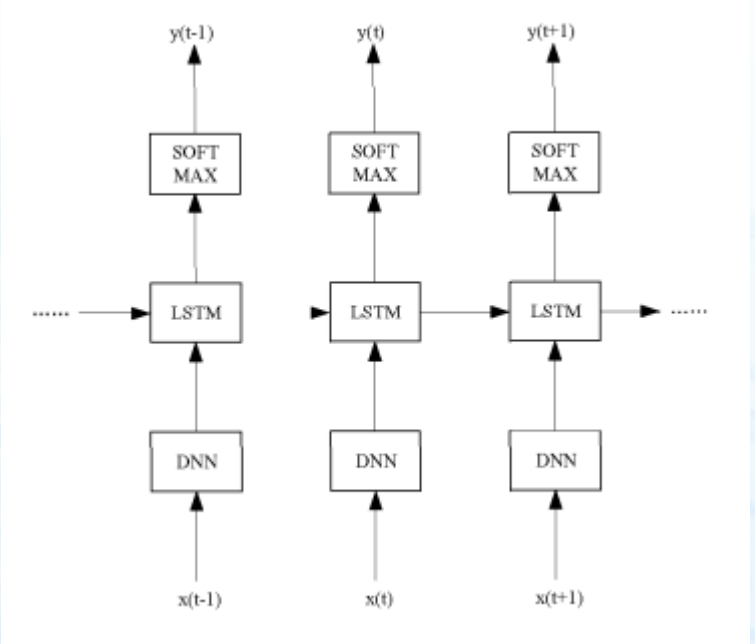
- 考虑到了相邻帧的信息对当前帧VAD判决的联系
- 它将当前帧与前后相邻帧的特征向量连接在一起，对应的判决标记也合并成一个向量连接起来，用于网络训练
- 最后一层网络输出被解释为对应的帧的语音/非语音概率 。最后，判决概率通过对 W 个输出取平均获得。根据 是否大于门 限，判断是否代表语音。

$$\hat{y}_L = \frac{\sum_{w=-(W-1)/2}^{(W+1)/2} y_L(w)}{W} \qquad \bar{y}_n = \begin{cases} 1 & if\ \hat{y}_n \geq \eta \\ 0 & otherwise \end{cases}$$

DNN-LSTM-VAD

特点：

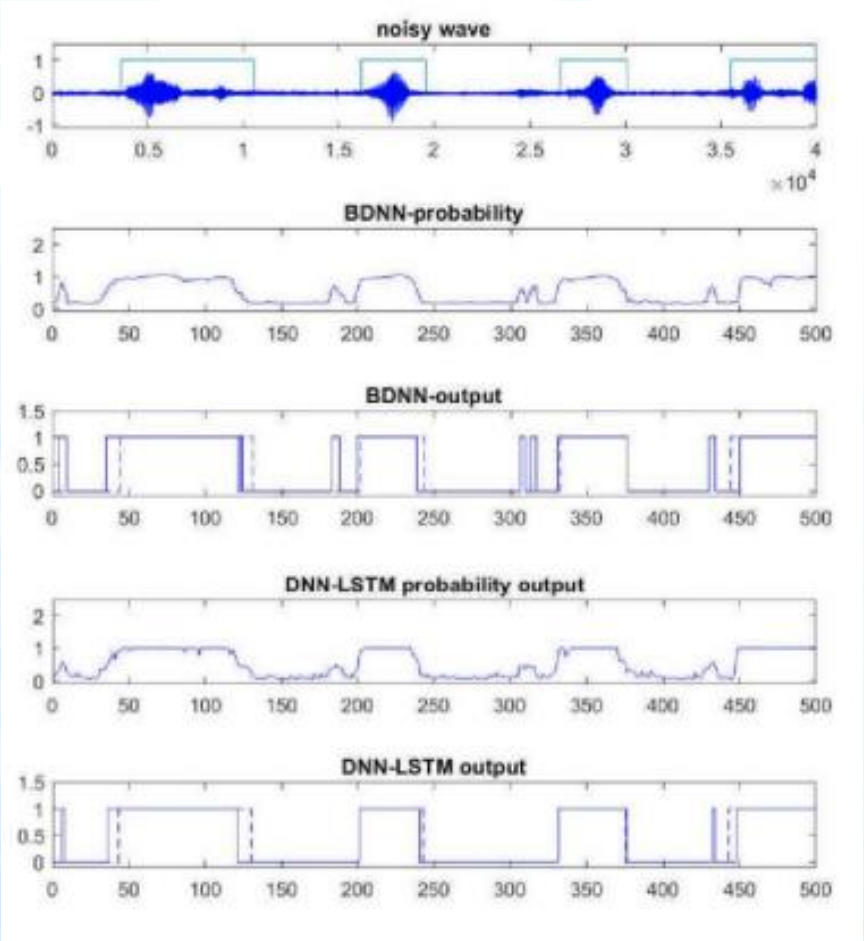
- 一个多层DNN加一层LSTM的网络结构
- 元素按时间顺序每个时刻通过 DNN层参与第 t时刻LSTM网络的计算。每一时刻的输出再通过进行语音/非语音后验概率的输出



BDNN与DNN-LSTM

1，DNN-VAD的输出在语音区内部更加平缓，但输出结果不稳定，会出现一些时间较长的突起，同时对语音的开始与结束阶段的检测不灵敏；

2，DNN-LSTM-VAD虽然在语音区内有一定的波动，但语音区的概率始终保持在 0.9以上，同时在分界点输出概率可以迅速地变化



采用基于上下文信息的训练

可以看出，相比于DNN-LSTM-VAD，尽管DNN-LSTM-VAD+在语音区内部有一定的波动，但幅度较大的突起数量明显减少，这可以使识别率有效提升。同时对标记的变化更加明显，延迟或提前的时间更短。