



中国科学院大学  
University of Chinese Academy of Sciences

# Deep Learning

## Application in Speech

---

Xinfeng Zhang (张新峰)

School of Computer Science and Technology

University of Chinese Academy of Sciences

Email: [xfzhang@ucas.ac.cn](mailto:xfzhang@ucas.ac.cn)



计算机科学与技术学院

SCHOOL OF COMPUTER SCIENCE AND TECHNOLOGY



## 提纲

---

- 语音技术概览
- 常见语音数据集
- 语音识别
- 声纹识别
- 语音合成
- 中英文术语对照



# 1

## 语音技术概览

# 语音的定义

- 语音指的是人们讲话时发出的话语
- 是组成语言的声音或者带有语言信息的声音
- 是一种人们进行信息交流产生的声音

语音(Speech)=声音(Acoustic) + 语言(Language)



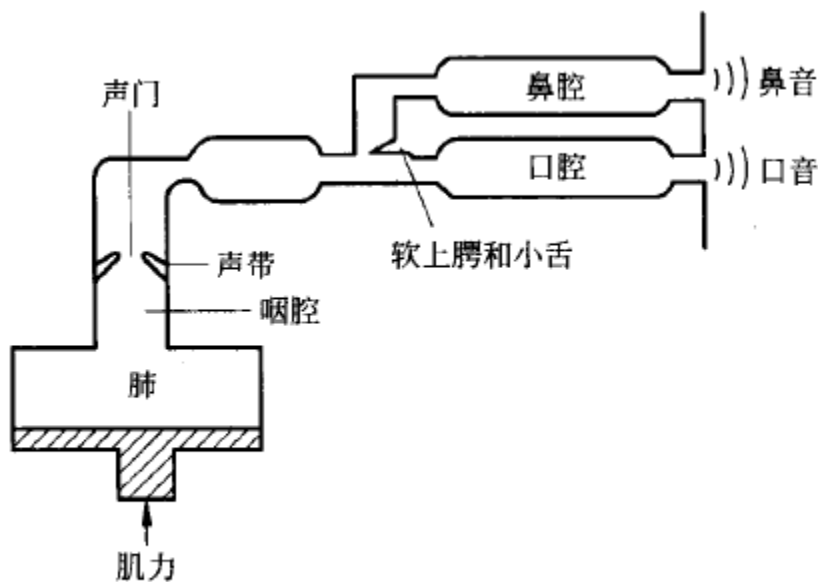
# 语音信号的产生

## □ 激励源：气流和声带

- 声带振动频率：基音频率
- 清音：声带不振动
- 浊音：声带振动

## □ 声道：可变谐振腔

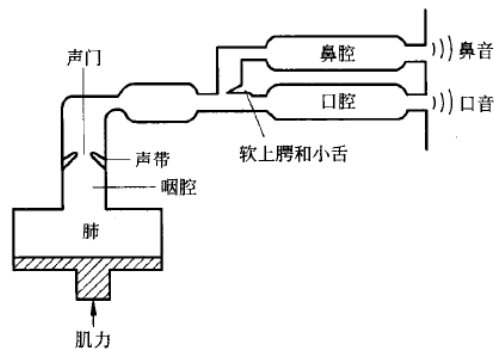
- 不同形状、不同声音
- 共振(谐振)频率



语音产生的机理图

# 发音的分类

- ❑ **浊音** (voiced sounds) : 声道打开, 声带先打开后关闭, 气流经过**使声带发生张弛振动**, 变为准周期振动气流。浊音的激励源被等效为准周期的脉冲信号
- ❑ **清音** (unvoiced sounds) : **声带不振动**, 而在声道某处保持收缩, 气流在声道里收缩后高速通过产生湍流, 再经过主声道 (咽、口腔) 的调整最终形成清音。清音的激励源被等效为一种白噪声信号
- ❑ **爆破音** (plosive sounds) : 声道关闭之后产生压缩空气然后突然打开声道所发出的声音



语音产生的机理图

# 语音的声学特性

- ❑ 音色: 又称为音质, 是一种声音区别于另一种声音的基本特性。与人声带的振动频率、发音器官的送气方式和声道的形状、尺寸密切相关
- ❑ 音调: 声音的高低, 取决于声波的频率
- ❑ 音强: 声音的强弱, 它由声波的振动幅度所决定
- ❑ 音长: 声音的长短, 取决于发音持续时间的长短



# 语音信号

## □ 语音信号的时域波形



(a) 语音信号“开始”时域波形



(b) 元音部分/ai/展开图



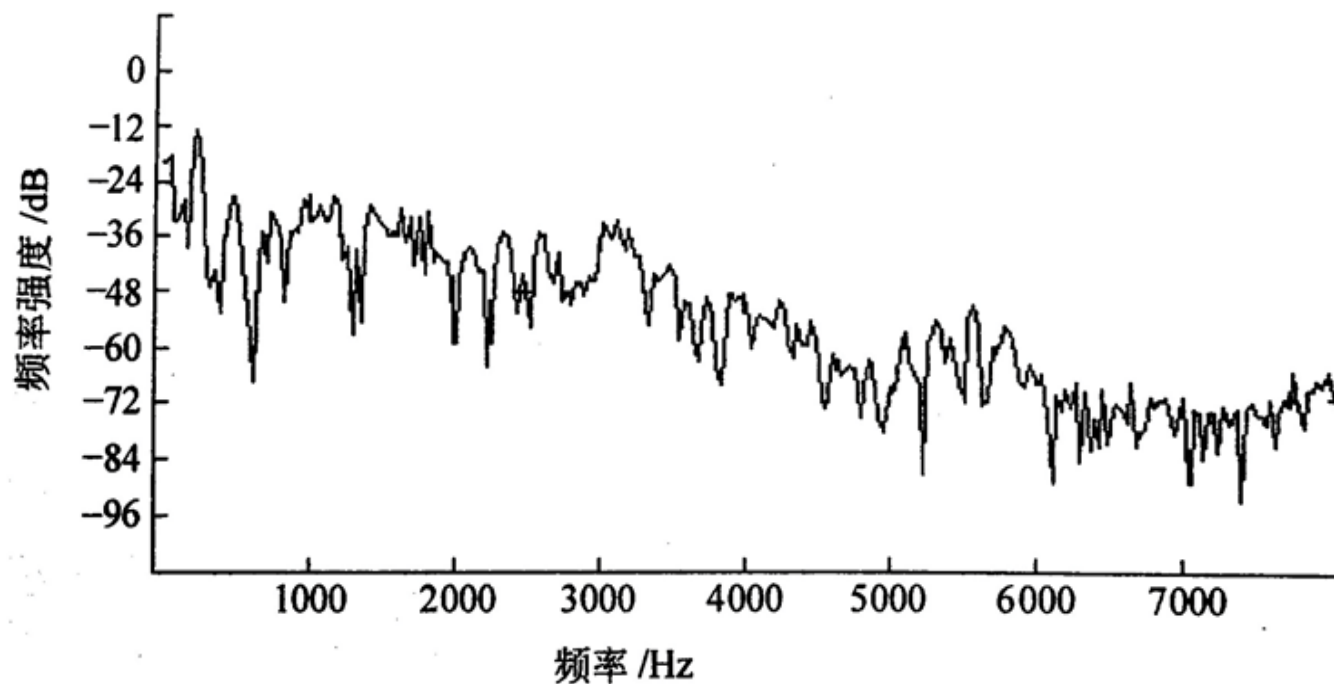
(c) 辅音部分/k/的展开图

语音信号“开始”的时域波形及其展开图



# 语音信号

## □ 语音信号的频域波形



“开始”中/ai/的频谱特性

# 语音信号处理

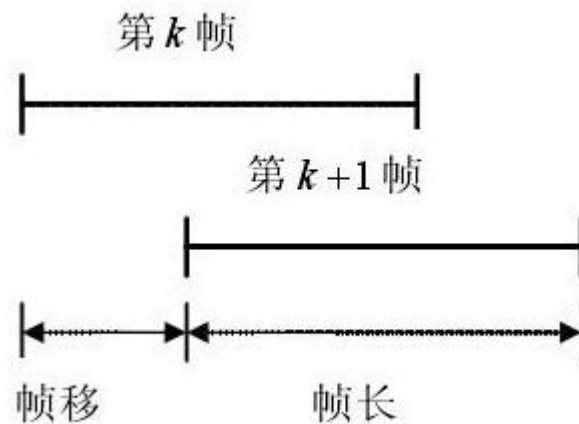
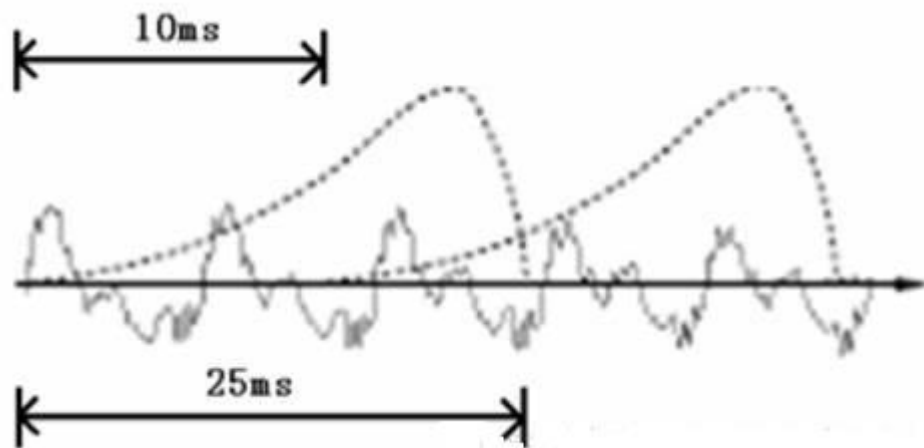
---

- 语音信号处理涉及语言学、声学、认知科学、生理学、心理学和数理统计等多学科知识
- 语音信号处理的目标就是使机器像人一样“能听会说”



# 语音信号处理

- ❑ 分帧：短时分析时将语音流分为一段一段来处理，每一段称为一“帧”
- ❑ 帧长：帧的时间跨度。10~30ms，常用20ms
- ❑ 帧移：帧与帧之间的平滑过度，0~1/2帧长



# 语音信号处理

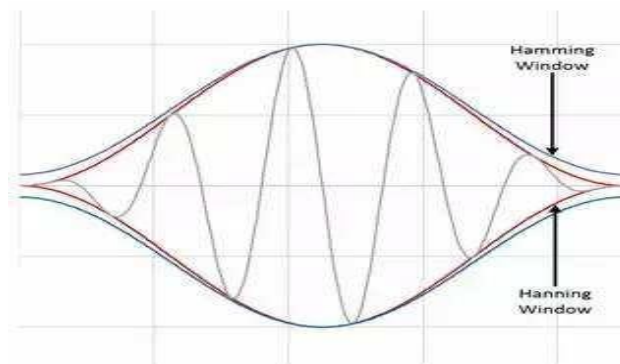
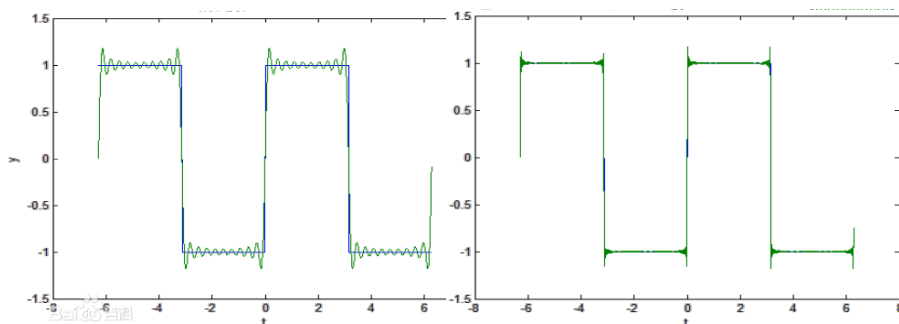
❑ 加窗：为防止**吉布斯（Gibbs）效应**，需要加窗处理

❑ 方法：矩形窗、Hamming、Hanning，通常采用**Hamming**

– 矩形窗： $w(n) = R_N(n) = 1 \quad n = 0, 1, \dots, N-1$

– Hamming： $w(n) = \left[ 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \right] R_N(n)$

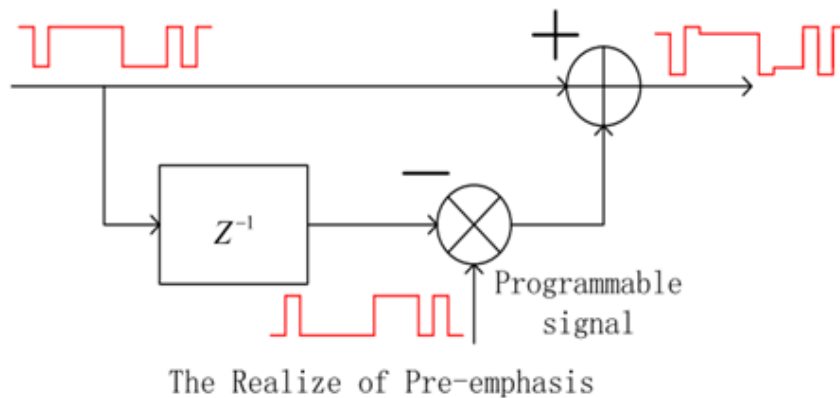
– Hanning： $w(n) = \frac{1}{2} \left[ 1 - \cos\left(\frac{2\pi n}{N-1}\right) \right] R_N(n)$



# 语音信号处理

- 预加重：预加重（Pre-emphasis）是一种在发送端事先对语音信号的高频分量进行补偿的方法，目的是减少尖锐噪声影响，提升高频部分

$$y[n] = x[n] - \alpha \cdot x[n-1] \quad 0.9 < \alpha < 1.0$$



# 语音信号处理

## □ 短时能量

- 通常指一帧语音段的能量
- 语音段的能量比噪声段的能量大
- 浊音的能量值比清音大得多

## □ 短时能量计算（加窗函数是矩形框时）

- 平方对数和  $E = \sum_{i=1}^N \log x(i)^2$
- 平方和  $E = \sum_{i=1}^N x(i)^2$
- 绝对值  $E = \sum_{i=1}^N |x(i)|$



# 语音信号处理

## □ 过零率

- 过零就是指信号通过零值
- 过零率就是每秒内信号值通过零值的次数
- 短时能量可以近似为互补的情况，短时能量大的地方过零率小，短时能量小的地方过零率较大

$$Z = \frac{1}{2} \left\{ \sum_{n=1}^{N-1} |\text{sgn}[s_w(n)] - \text{sgn}[s_w(n-1)]| \right\}$$




# 语音信号处理

## □ 线性预测模型（Linear Prediction Coefficients, LPC）

- 一个语音的抽样能够用过去若干个语音抽样的线性组合来逼近
- 这个线性预测的抽样和实际语音抽样之间存在着误差
- 通过实现预测采样在最小均方误差意义上逼近实际采样，可以求取一组唯一的预测系数

预测系数

$$\hat{x}(n) = \sum_{i=1}^p a_i x(n-i)$$


$$e(n) = x(n) - \hat{x}(n)$$

$$E = \sum_n e^2(n) = \sum_n [x(n) - \hat{x}(n)]^2$$





# 语音信号处理

## □ Linear Prediction Cepstral Coefficients (LPCC)

- 是LPC在倒谱域中的表示
  - 语音信号的倒谱可以通过对信号做傅里叶变换，取模的对数，再求反傅里叶变换得到
- 用多个倒谱系数可以代表共振峰的特性
- 在语音识别中取得很好的性能

$$\begin{cases} \hat{h}(0) = 0 \\ \hat{h}(1) = a_1 \\ \hat{h}(n) = a_n + \sum_{k=1}^{n-1} (1 - k/n) a_k \hat{h}(n-k) & 1 \leq n \leq p \\ \hat{h}(n) = \sum_{k=1}^p (1 - k/n) a_k \hat{h}(n-k) & n > p \end{cases}$$



# 语音信号处理

## □ 梅尔频率倒谱系数特征提取

- Mel-Frequency Cepstral Coefficients (MFCC)
- 信号的预处理，包括预加重(Pre-emphasis)，分帧(Frame Blocking)，加窗(Windowing)。假设语音信号的采样频率 $f_s=8\text{KHz}$ ，由于语音信号在10-30ms认为是稳定的，则可设置帧长为80~240点。帧移可以设置为帧长的1/2
- 对每一帧进行FFT变换，求频谱，进而求得幅度谱



# 语音信号处理

## □ 梅尔频率倒谱系数特征提取

- 对幅度谱加Mel滤波器组

$$B(f) = 1125 \ln(1 + f/700) \quad f \text{ -- 频率} \quad B \text{ -- Mel-频率}$$

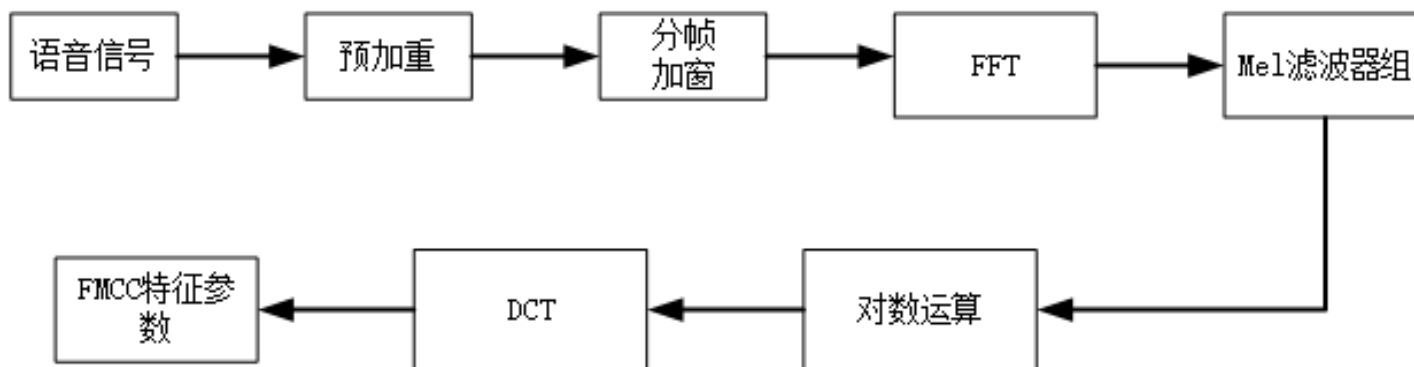
- 对所有的滤波器输出做对数运算(Logarithm), 再进一步做离散余弦变换(DCT) 可得MFCC

$$c(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^L \log m(l) \cos \left\{ \left( l - \frac{1}{2} \right) \frac{i\pi}{L} \right\} \quad L \text{ 为滤波器个数}$$



# 语音信号处理

## □ 梅尔频率倒谱系数特征提取



# 语音信号的采集和存储

## □ 语音信号的采集:

- 可以使用Windows系统自带的“录音机”进行录音。如果有更高要求，需使用专用设备

## □ 语音信号的存储:

- 波形音频文件：一种最直接的表达声波的数字形式，“.wav”
- MIDI音频文件：计算机数字音乐接口生成的音频文件，“.mid”
- 压缩音频文件：一种MP3格式的压缩音频文件，“.mp3”
- .....



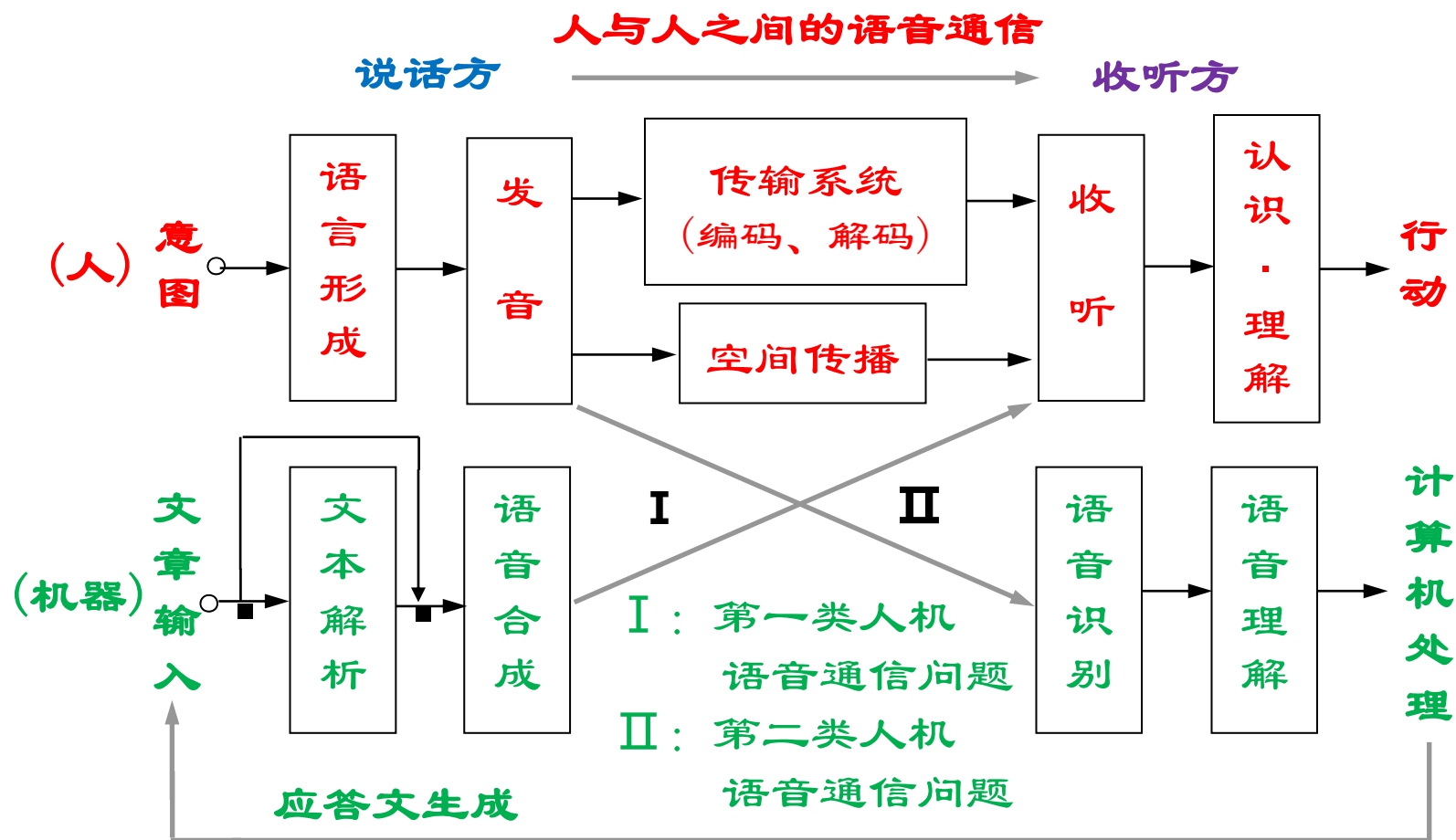
# 语音信号处理的主要类别

---

- ❑ 语音识别：识别表达语言的语音内容
- ❑ 声纹识别：识别特定语音对应的人
- ❑ 语音合成：将文本转换成相应的语音



# 人与人、人与机器间语音信息处理过程



# 语音技术发展史

---

- ❑ 50年代：AT&T Bell Lab，可识别10个英文数字
- ❑ 60年代：线性预测编码（Linear Prediction Coefficient, LPC）较好地解决了语音信号产生模型，动态规划（Dynamic Programming, DP）则有效解决了不等长语音的匹配问题
- ❑ 70年代：动态时间规整（Dynamic Time Warp, DTW）技术基本成熟，实现了基于LPC和DTW技术相结合的特定人孤立词语音识别系统





# 语音技术发展史

---

- ❑ 80年代：HMM模型和人工神经网络（ANN）在语音识别中成功应用。1988年美国CMU大学基于HMM开发SI-CSR系统 SPHINX
- ❑ 90年代：大规模应用，理论进展缓慢
- ❑ 2001年：语音识别达到了80%的准确度，但此后鲜有进展
- ❑ 2010年：深度学习方法的使用，语音识别取得突破性进展



# 语音技术的典型应用





# 2

## 常见语音数据集

# 常见语音数据集

## □ THCHS30

- 由清华大学语音与语言技术中心（CSLT）出版的开放式免费中文语音数据库
- 包含了1万余条语音文件，大约40小时的中文语音数据，内容以文章诗句为主，全部为女声
- 数据库对学术用户完全免费
- <https://arxiv.org/abs/1512.01882>
- <https://www.openslr.org/18/>

### THCHS-30

**Identifier:** SLR18

**Summary:** A Free Chinese Speech Corpus Released by CSLT@Tsinghua University

**Category:** Speech

**License:** Apache License v.2.0

**Downloads (use a mirror closer to you):**

[data\\_thchs30.tgz](#) [6.4G] ( speech data and transcripts ) Mirrors: [\[China\]](#)

[test-noise.tgz](#) [1.9G] ( standard 0db noisy test data ) Mirrors: [\[China\]](#)

[resource.tgz](#) [24M] ( supplementary resources, incl. lexicon for training data, noise samples ) Mirrors: [\[China\]](#)

# 常见语音数据集

## □ AISHELL

- 由北京希尔公司发布的一个免费中文语音数据集
- 包含约178小时的开源版数据
- 该数据集包含400个来自中国不同地区、具有不同的口音的人的语音
- 该数据免费供学术使用
- <https://arxiv.org/abs/1709.05522>
- <https://www.openslr.org/33/>

### Aishell

**Identifier:** SLR33

**Summary:** Mandarin data, provided by Beijing Shell Shell Technology Co.,Ltd

**Category:** Speech

**License:** Apache License v.2.0

**Downloads (use a mirror closer to you):**

[data\\_aishell.tgz](#) [15G] ( speech data and transcripts ) Mirrors: [\[China\]](#)

[resource\\_aishell.tgz](#) [1.2M] ( supplementary resources, incl. lexicon, speaker info ) Mirrors: [\[China\]](#)

# 常见语音数据集

## □ ST-CMDS

- 由一个AI数据公司发布的免费中文语音数据集。
- 包含10万余条语音文件，大约100余小时的语音数据。
- 数据内容以平时的网上语音聊天和智能语音控制语句为主，855个不同说话者，同时有男声和女声
- <https://www.openslr.org/38/>

## Free ST Chinese Mandarin Corpus

**Identifier:** SLR38

**Summary:** A free Chinese Mandarin corpus by Surfingtech (www.surfing.ai), containing utterances from 855 speakers, 102600 utterances;

**Category:** Speech

**License:** Creative Common BY-NC-ND 4.0 (Attribution-NonCommercial-NoDerivatives 4.0 International)

**Download:** [ST-CMDS-20170001\\_1-OS.tar.gz](#) [8.2G] ( speech audios and transcripts ) Mirrors: [\[China\]](#)

**About this resource:**

# 常见语音数据集

## □ Primewords Chinese Corpus Set 1

- 由上海普力信息技术有限公司发布的免费中文普通话语料库。
- 包含了大约100小时的中文语音数据，语料库由296名母语为中文的智能手机录制。
- 学术用途免费
- <https://www.openslr.org/47/>

## Primewords Chinese Corpus Set 1

**Identifier:** SLR47

**Summary:** Chinese Mandarin corpus released by Shanghai Primewords Co. Ltd. ([www.primewords.cn](http://www.primewords.cn)), containing 100 hours of speech data.

**Category:** Speech

**License:** Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

**Download:** [primewords\\_md\\_2018\\_set1.tar.gz](#) [9.0G] (speech data and transcripts) Mirrors: [\[China\]](#)

# 常见语音数据集

## □ TIMIT

- 由德州仪器、麻省理工学院和SRI International合作构建的声学—音素连续语音语料库。
- TIMIT数据集的语音采样频率为16kHz，一共包含6300个句子
- 语音由来自美国八个主要方言地区的630个人每人说出给定的10个句子，所有的句子都在音素级别（phone level）上进行了手动分割，标记
- <https://catalog.ldc.upenn.edu/LDC93S1>

Home > Language Resources > Data

### TIMIT Acoustic-Phonetic Continuous Speech Corpus

Item Name:	TIMIT Acoustic-Phonetic Continuous Speech Corpus
Author(s):	John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue
LDC Catalog No.:	LDC93S1
ISBN:	1-58563-019-5
ISLRN:	664-033-662-630-6
Member Year(s):	1993
DCMI Type(s):	Sound
Sample Type:	1-channel pcm
Sample Rate:	16000
Data Source(s):	microphone speech
Application(s):	speech recognition
Language(s):	English
Language ID(s):	eng
License(s):	<a href="#">LDC User Agreement for Non-Members</a>
Online Documentation:	<a href="#">LDC93S1 Documents</a>
Licensing Instructions:	<a href="#">Subscription &amp; Standard Members, and Non-Members</a>
Citation:	Garofolo, John S., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.



# 常见语音数据集

---

## □ TED-LIUM Corpus

- 包括TED演讲音频和对应讲稿。其中包括1495段演讲录音和对应的演讲稿，数据获取自TED网站
- <https://www.openslr.org/51/>

## □ VoxForge

- 该数据集是带口音的语音清洁数据集，对测试模型在不同重音或语调下的鲁棒性非常有用
- <http://www.voxforge.org/>



3

语音识别

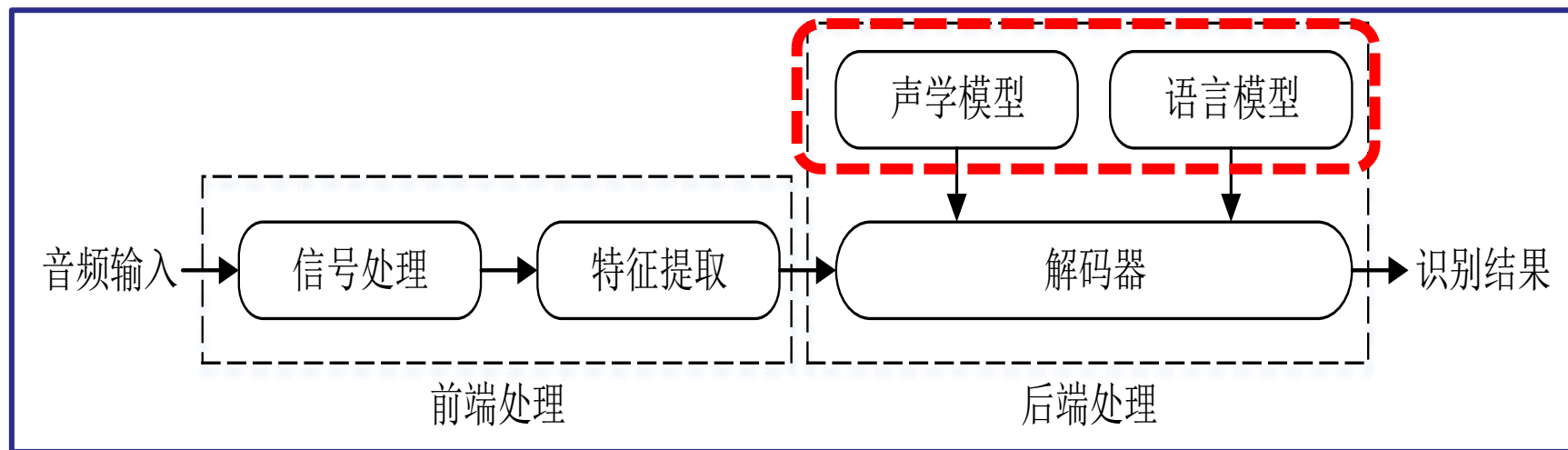
# 语音识别基本概念

- 语音识别(Speech Recognition, SR)是以语音信号为研究对象，让机器通过识别和理解的过程，将**语音信号转为相应文字或命令**的技术
- 目的是**让机器“听懂”人说话**，是人机交互的重要方式之一



# 语音识别基本概念

## □ 技术框架



- 声学模型（Acoustic Model, AM）的任务是建模给定文本下产生语音波形的概率
  - 将声学 and 发音学的知识进行整合，以特征提取模块提取的特征为输入，生成声学模型得分
  - 声学模型是语音识别系统的重要组成部分，它占据着语音识别大部分的计算开销，决定着语音识别系统的性能

# 声学模型：GMM-HMM

---

- ❑ 高斯混合模型（Gaussian mixture model, GMM）用于对语音信号的**声学特征分布**进行建模
- ❑ 隐马尔科夫模型（Hidden Markov model, HMM）则用于对**语音信号的时序性**进行建模
- ❑ 维特比算法（Viterbi）：针对**篱笆网络的有向图（Lattice）的最短路径问题而提出的动态规划算法**。凡是使用隐含马尔可夫模型描述的问题都可以用维特比算法来解码

# 声学模型：GMM-HMM

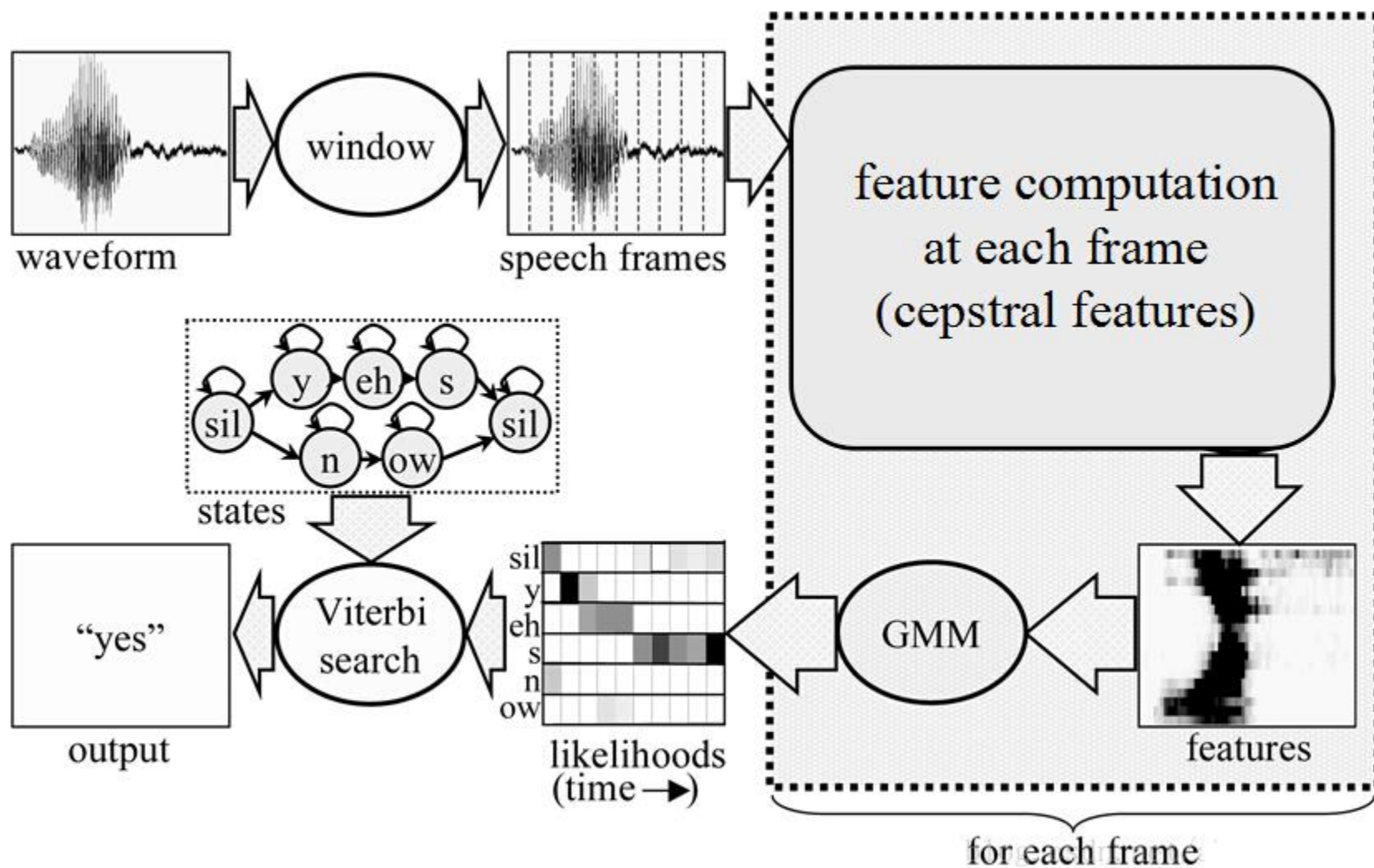
---

## □ GMM-HMM语音识别分三步：

- 第一步，把帧识别成状态（难点），GMM
- 第二步，把状态组合成音素，HMM
- 第三步，把音素组合成单词，HMM

# 声学模型：GMM-HMM

## □ 框图



# 声学模型： DNN-HMM

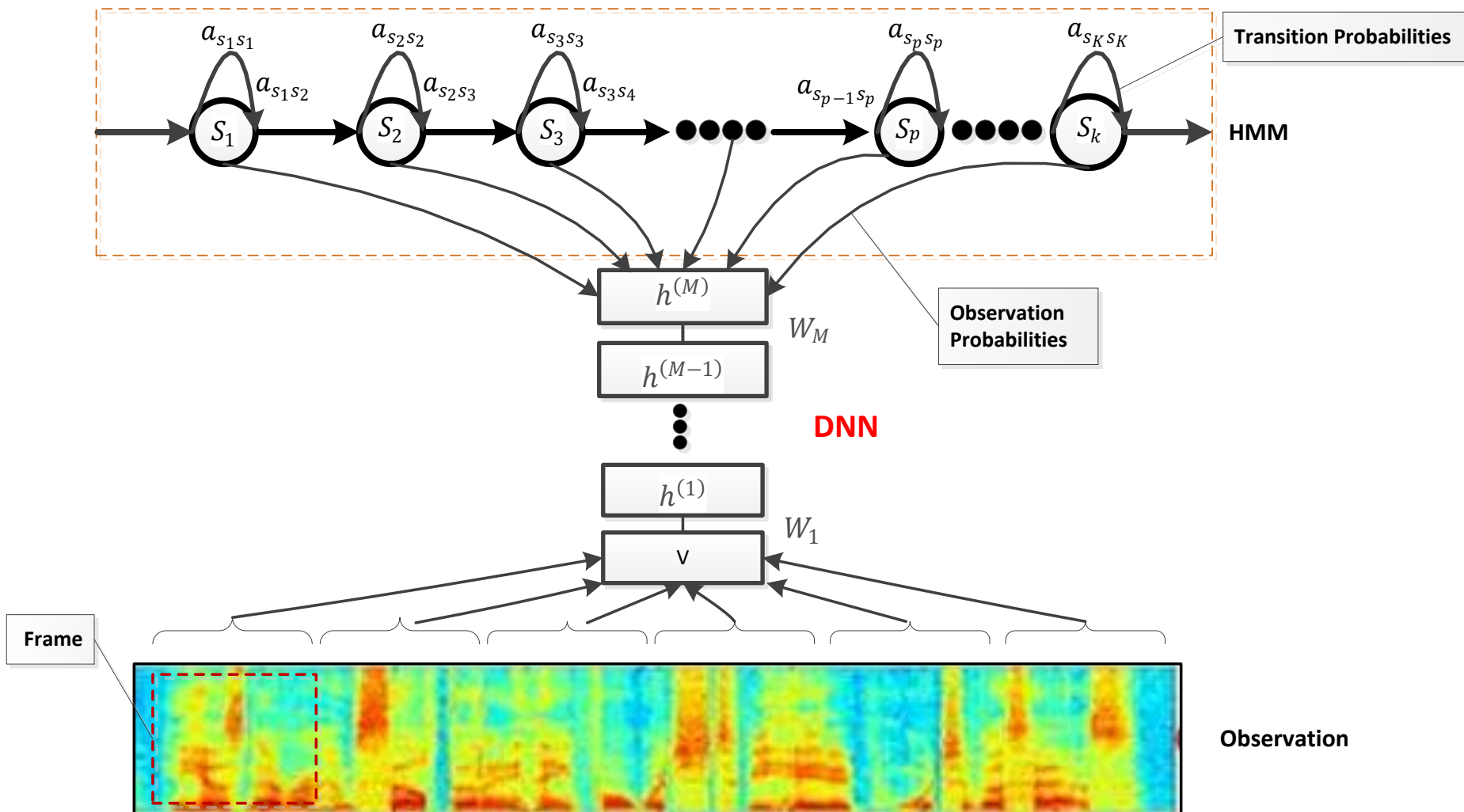
---

- ❑ GMM模拟任意函数的功能取决于混合高斯函数的个数，所以具有一定的局限性，属于浅层模型
- ❑ 深度神经网络可以模拟任意的函数，因而表达能力更强
- ❑ 随着深度学习的发展，DNN模型展现出了明显超越GMM模型的性能，于是替代了GMM进行HMM状态建模



# 声学模型：DNN-HMM

## □ 框图



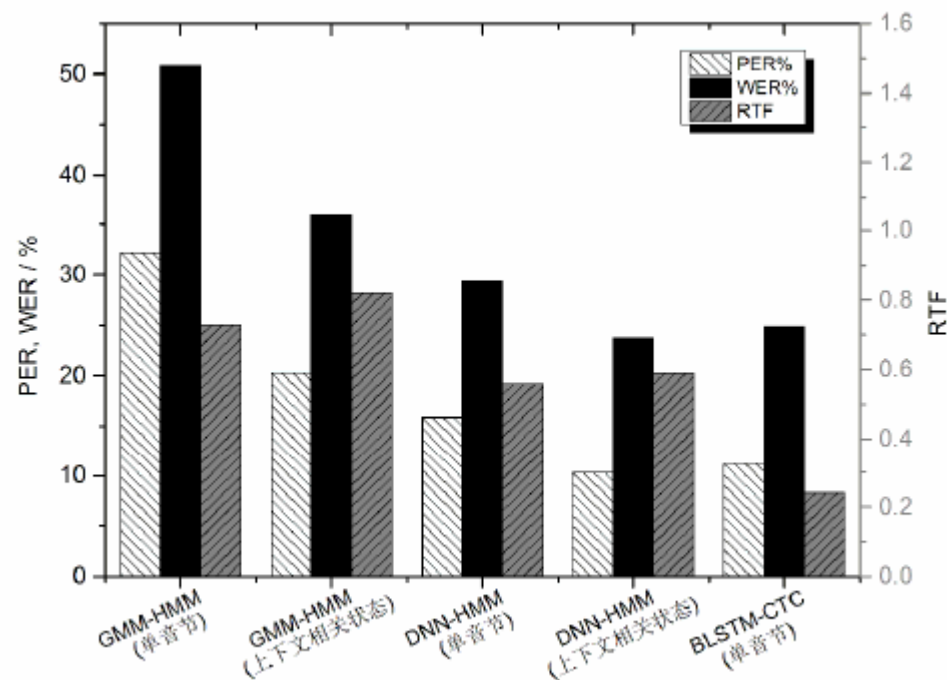
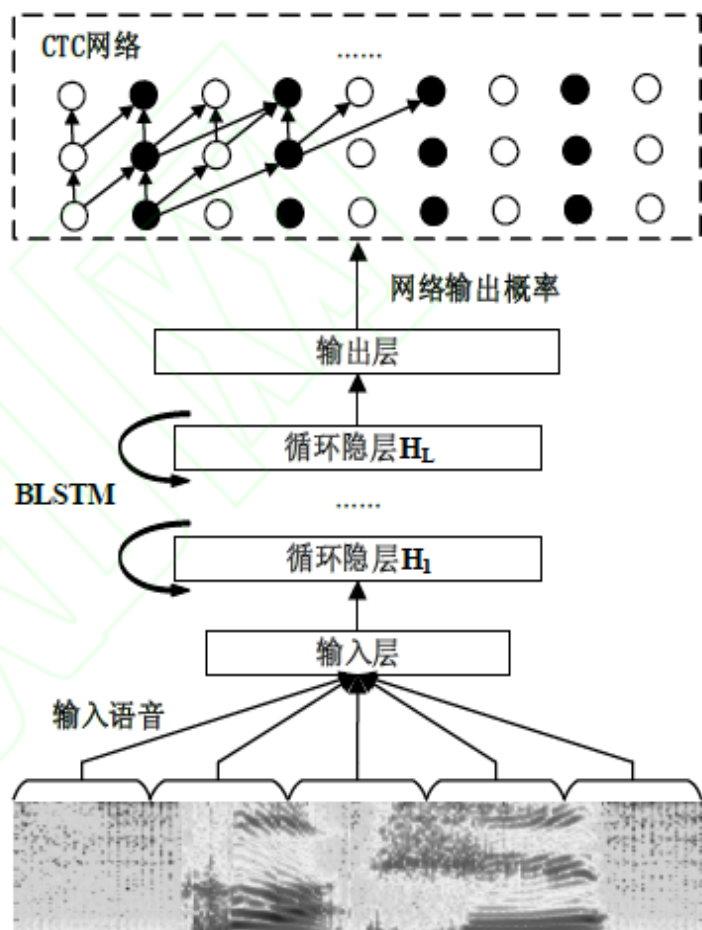
# 声学模型：BLSTM-CTC

---

- ❑ 然而在混合DNN/HMM系统的训练过程中，依然**需要利用GMM来对训练数据进行强制对齐**，以获得语音帧层面的标注信息进一步训练DNN。这样显然**不利于针对整句发音进行全局优化**，同时也相应地增加了识别系统的复杂度和搭建门槛
- ❑ 对于序列标记任务，Graves等人提出了在循环神经网络训练中**引入联结时序分类(Connectionist Temporal Classification, CTC)**目标函数，使得RNN可以**自动地完成序列输入自动对齐任务**，进而提出了**BLSTM-CTC模型**

# 声学模型：BLSTM-CTC

## □ 框图



PER (Phone Error Rate, 音节错误率)

WER (Word Error Rate, 词错误率)

RTF (Real Time Factor, 实时率)

SER (Sentence Error Rate, 句子错误率)

CER (Character Error Rate, 字错误率)

# 声学模型：DFCNN-CTC

---

- ❑ 深度全序列卷积神经网络（Deep Fully Convolutional Neural Network, DFCNN）：由科大讯飞2016年提出的一种使用深度卷积神经网络来对语音时频图进行识别的方法
- ❑ 连接时序分类（Connectionist temporal classification, CTC）：CTC不需要标签在时间上一一对齐就可以进行训练，在对输入数据的任一时刻做出的预测不是很关心，**而关心的是整体上输出是否与标签一致，从而减少了标签预划定的冗杂工作。**在整个网络结构中把CTC作为损失函数

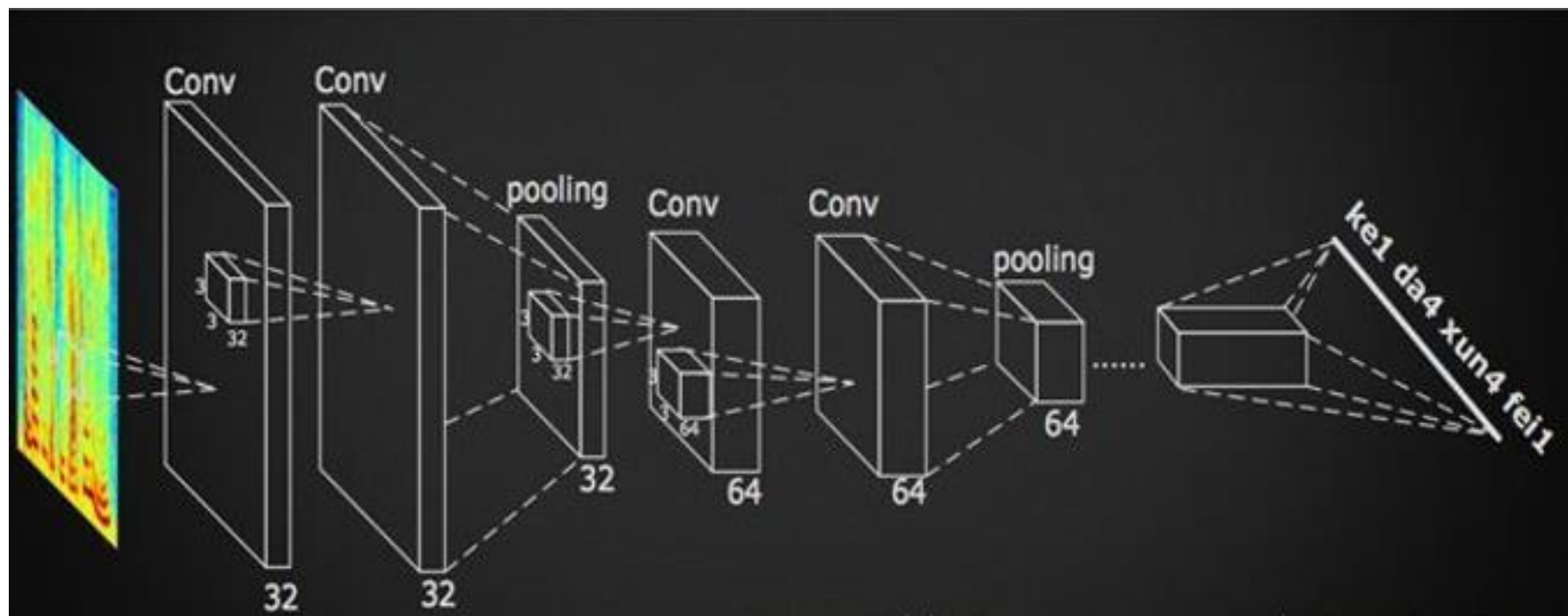
# 声学模型：DFCNN-CTC

---

- DFCNN 比较灵活，可以方便地和其他建模方式融合，比如和**连接时序分类模型(CTC)方案结合**，以实现整个模型的端到端声学模型训练
- 和目前（2016年）业界最好的语音识别框架BLSTM-CTC系统相比，DFCNN 系统获得了额外15%的性能提升

# 声学模型：DFCNN-CTC

- DFCNN 先对时域的语音信号进行傅里叶变换得到语音的语谱图，直接将一句语音转化成为一张图像作为输入，输出单元则直接与最终的识别结果（比如音节或者汉字）相对应



# 语音识别的主要应用

字幕生成



智能音箱



语音输入



# 语音识别的主要应用

---

## □ 智能家居

- 用语音可以控制电视机、VCD、空调、电扇、窗帘的操作

## □ 语音搜索

- 搜索内容直接以语音的方式输入，响应速度更快，适用于音乐、电影、小说等内容搜索场景，让搜索内容输入更加便捷，高效



# 语音识别的主要应用

---

## □ 人机对话

- 将语音识别为文字，毫秒级响应，可用于聊天机器人、故事机等近场语音识别环境，让人机对话更加流畅自然

## □ 语音输入

- 通过语音识别将语音转换为文字实现输入，如语音输入法等



# 4

## 声纹识别

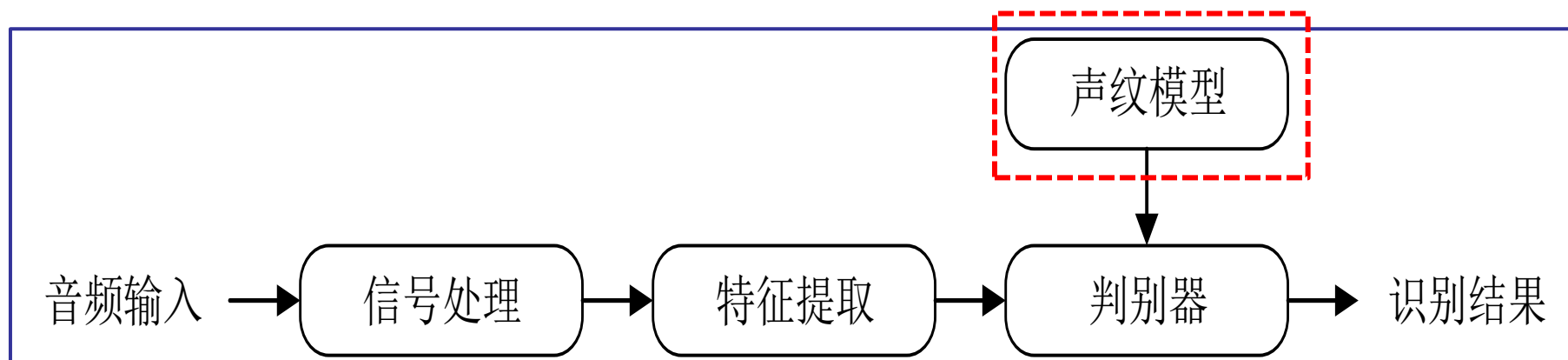


# 声纹识别的基本概念

- ❑ **声纹识别**（Voice Print Recognition, VPR），作为生物识别的一种，是根据说话人的声波特性进行身份辨识的服务
- ❑ 身份辨识与口音无关，与语言无关，可以用于**说话人辨认**和**说话人确认**
- ❑ 根据是否与说话内容有关，声纹识别又可分为：文本相关的声纹识别（Text-Dependent）、文本独立的声纹识别（Text-Independent）



# 声纹识别的技术框架



# 声纹模型：GMM-UBM

- 说话人识别最主要的两部分是特征提取和模式匹配，在模式匹配中，常用GMM
- 通用背景模型(Universal Background Model, UBM)描述的是语音特征在空间中的平均分布，且语音特征与目标说话者无关，与环境噪声和声道有关



# 声纹模型：GMM-UBM

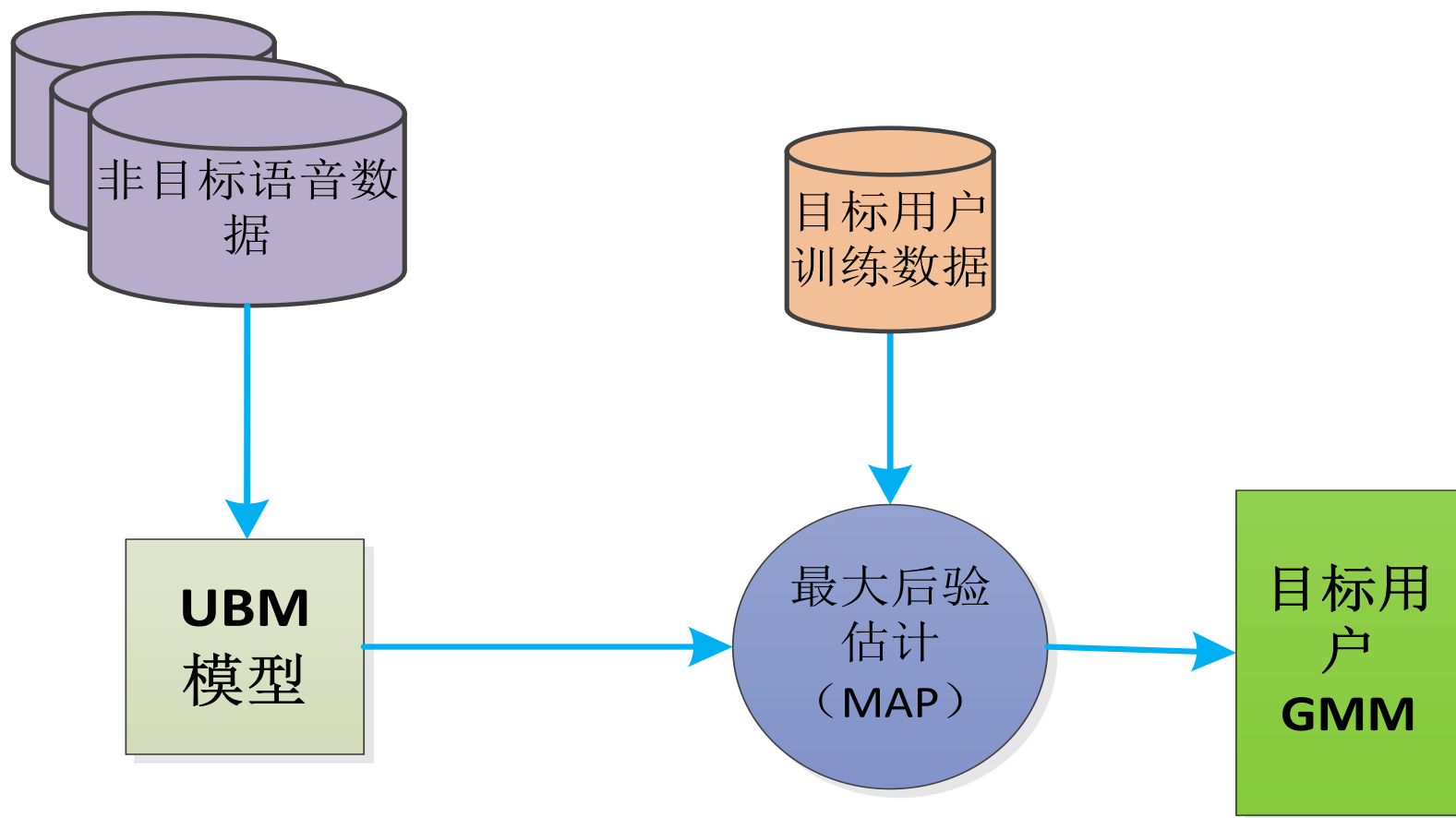
## □ 模型的流程

- 先使用大量的非目标用户数据训练UBM，然后使用极大后验概率(MAP)自适应算法和目标说话人数据来更新局部参数得到对应的GMM
- MAP自适应算法相当于先进行一轮EM迭代得到新的参数，然后将新参数和旧参数整合



# 声纹模型：GMM-UBM

模型框图



# 声纹模型：GMM-SVM

---

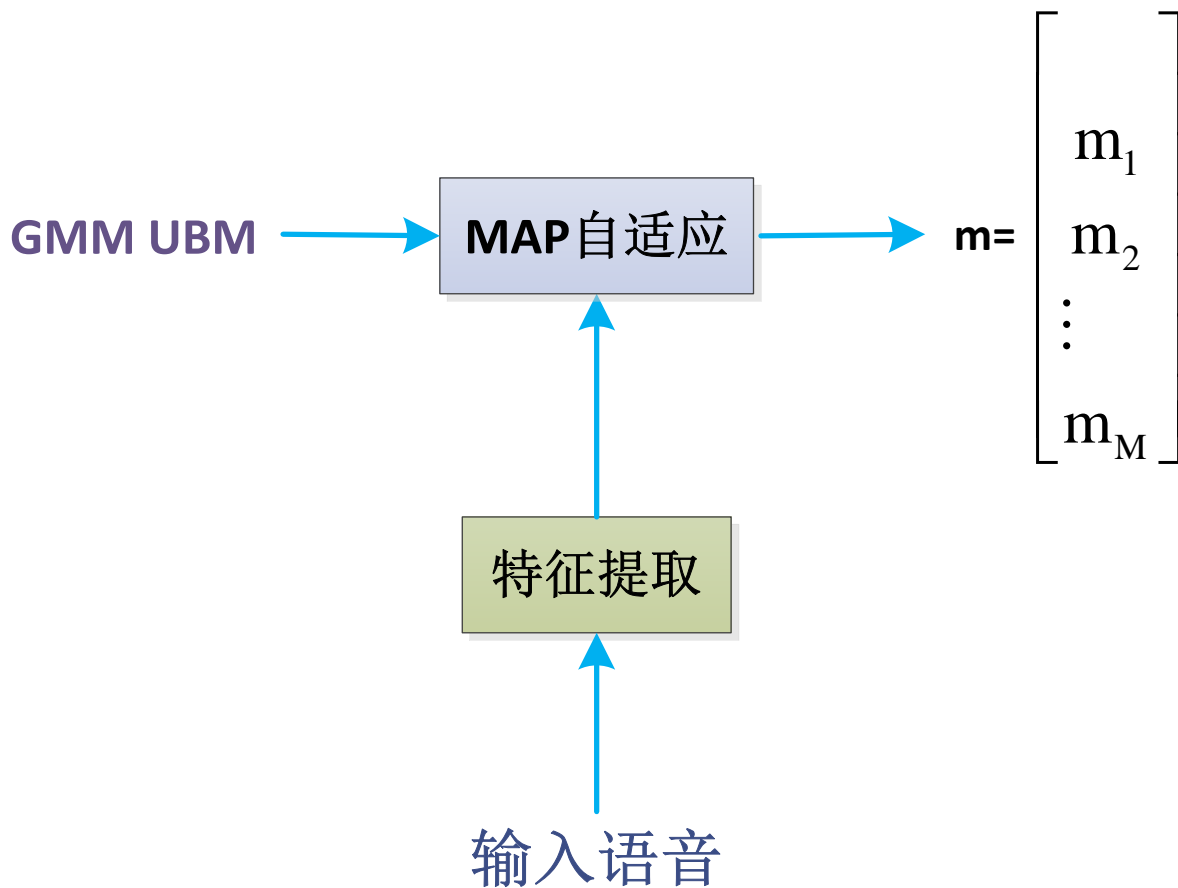
- 说话人识别：该模型对GMM中每个高斯分量均值构建一个**高斯超向量**（Gaussian Super Vector, GSV）作为SVM的样本
- 利用带核函数的SVM的非线性分类能力，在原始GMM-UBM的基础上大幅提升了识别性能





# 声纹模型：GMM-SVM

## GMM-Supervector的提取流程



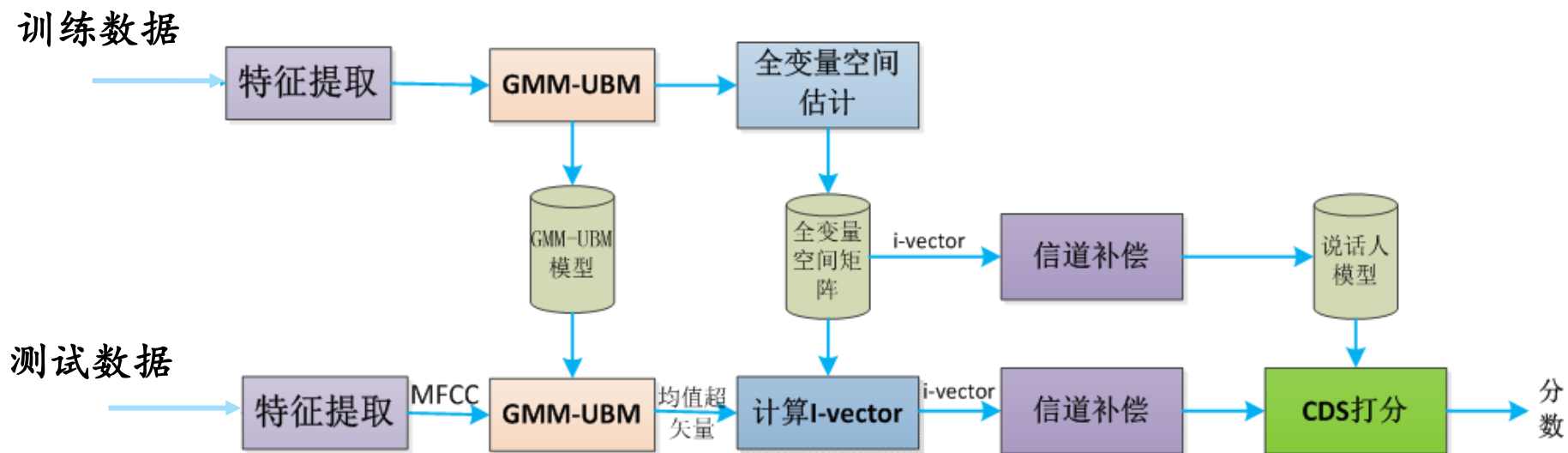
# 声纹模型：GMM-I-Vector

- ❑ Dehak提出了从GMM均值超向量中提取一个更紧凑的向量，称为I-Vector (Identity-Vector)
- ❑ Dehak提出了全局差异空间模型，将说话人差异和信道差异作为一个整体进行建模
- ❑ 当前，I-Vector在大多数情况下仍然是文本无关声纹识别中表现性能比较好的建模框架



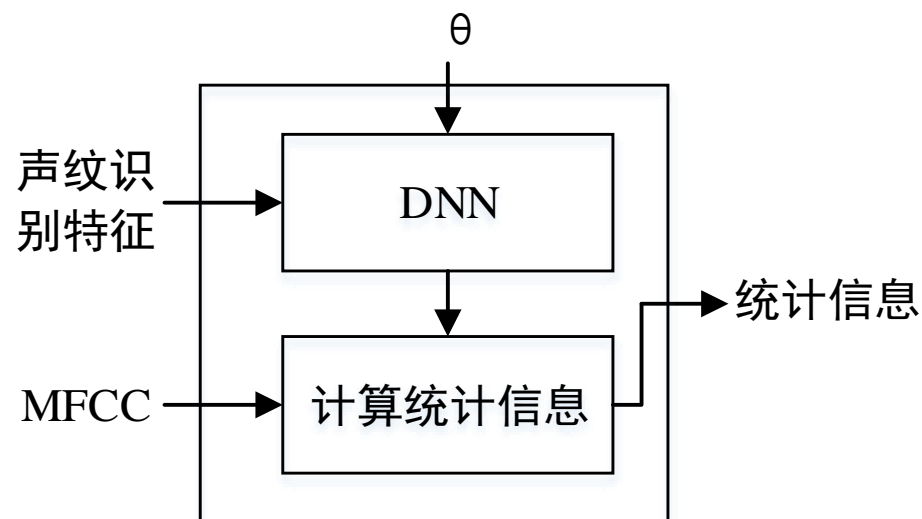
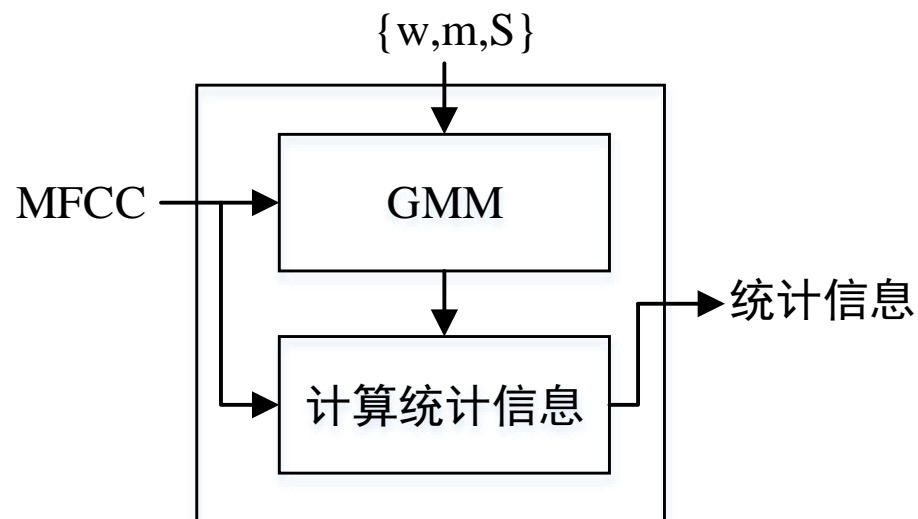
# 声纹模型：GMM-I-Vector

基于I-Vector的说话人识别框图



# 声纹模型：深度神经网络模型

## □ 传统模型上进行改进



- 梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)



# 声纹识别的主要应用

---

## □ 公共安全领域的声纹识别技术

- 公安司法人员还可以利用电话敲诈勒索等刑事案件的声音，绑架等方式识别技术，锁定嫌疑人的通话声，缩小刑事侦查范围



# 声纹识别的主要应用

---

## □ 金融身份认证

- 为了防止被盗刷子和其他情况的发生，将声纹确认技术添加到交易支付中，并通过动态声纹密码验证客户端语音身份，可以有效提高个人资金和交易支付的安全性
- 在国外，巴克莱银行，花旗银行，澳大利亚国家银行和万事达卡机构已开始引入声纹技术



# 声纹识别的主要应用

---

## □ 融合声纹技术的个性化的语音互动时代

- 利用声纹辨认技术，可支持智能音箱、智能语音助手等提供个性化服务，如针对家庭用户中的老年人、儿童等不同年龄段用户，按照兴趣推荐不同的歌曲、新闻等

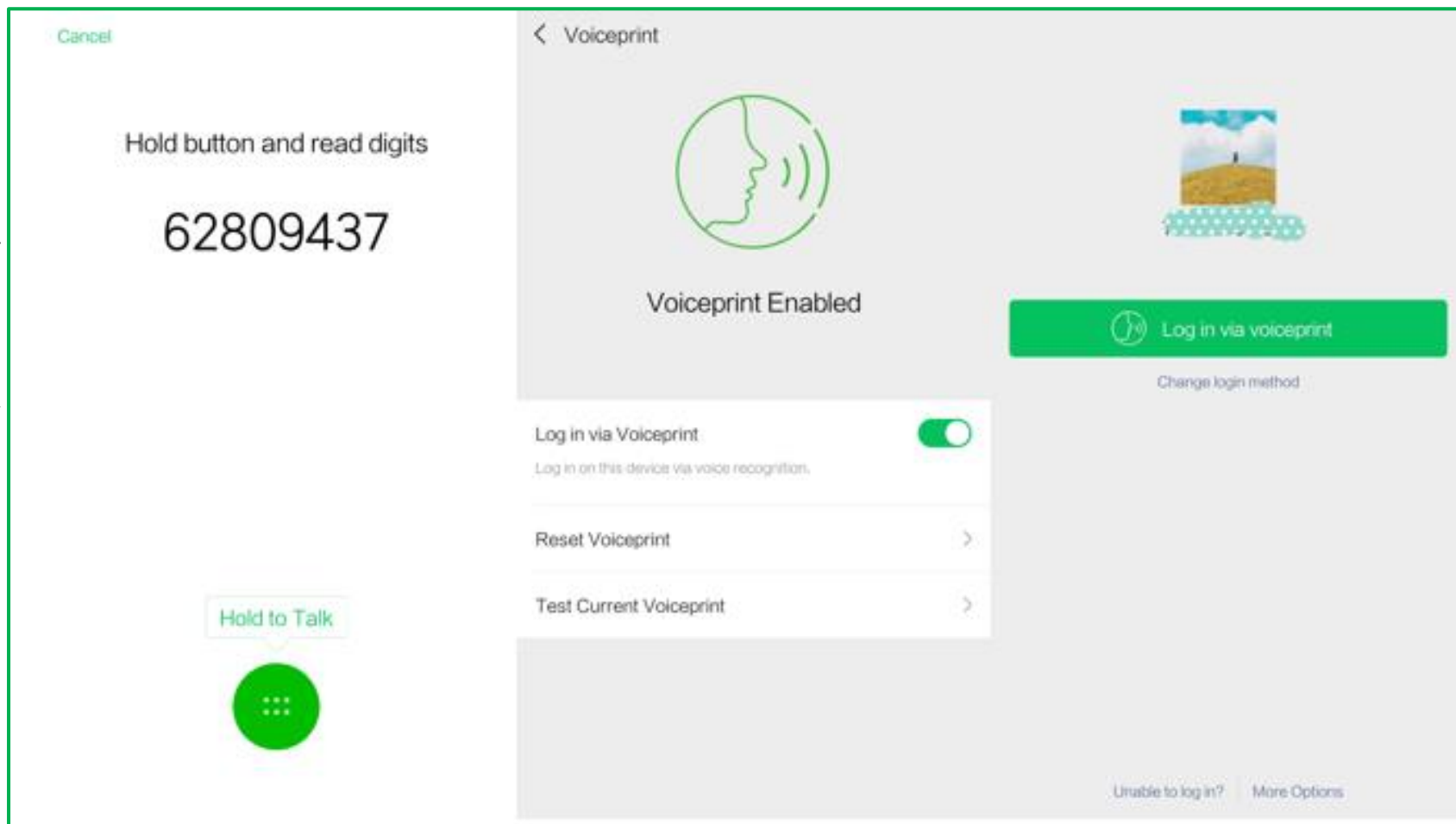
## □ 设备的访问控制授权

- 比如智能手机锁屏、各类网络账号的声控密码锁、电脑声控锁、声控安全门、汽车声控锁等



# 声纹识别的应用案例

微信声纹登录功能







# 5

## 语音合成



# 语音合成的基本概念

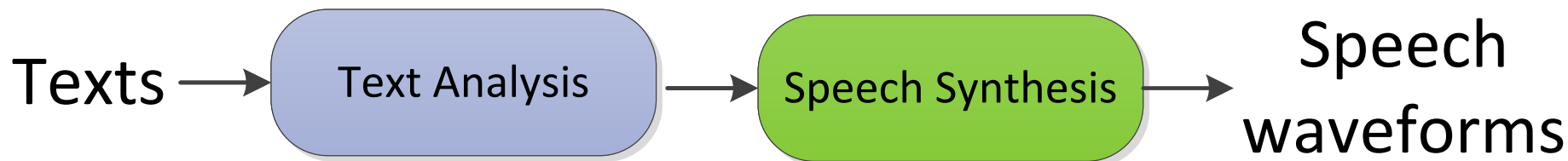
---

- 语音合成 (Text To Speech, TTS) 是将文本转为语音的技术
- 语音合成是实现人机语音交互, 建立一个有听和讲能力的交互系统所必需的关键技术



# 语音合成的基本框架

---



# 语音合成模型：WAVENET

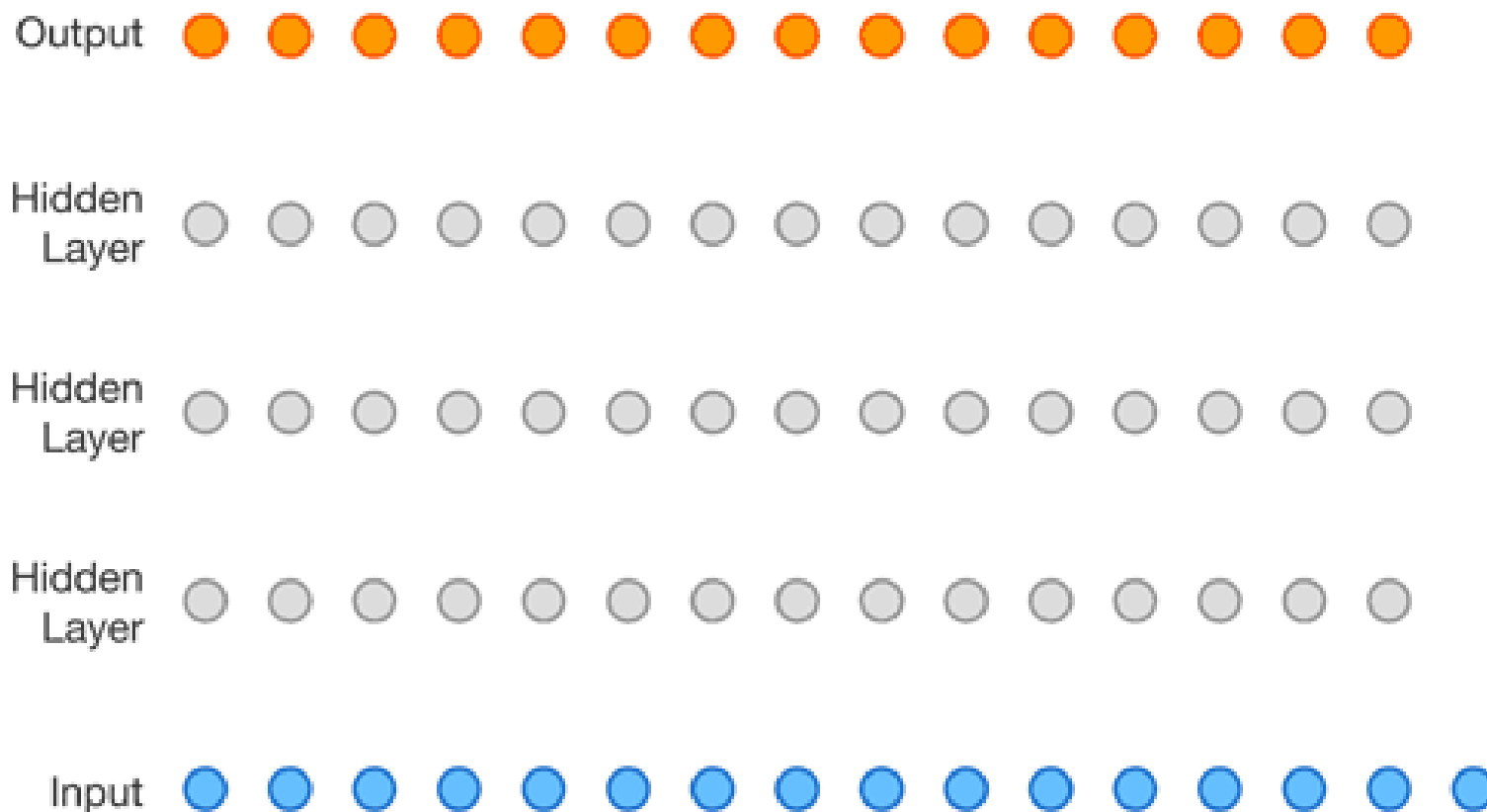
---

- ❑ **WaveNet**模型是一种序列生成模型，可以用于语音生成建模
- ❑ 2017年由DeepMind提出，在TTS(文字转语音)任务上可以达到当时state-of-art的效果
- ❑ 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



# 语音合成模型：WAVENET

## □ WaveNet动态展示



# 语音合成模型：Parallel WaveNet

---

- DeepMind公司2017年对此前的WaveNet版本进行改进
- 使用一个经过完全训练的 WaveNet 模型作为“教师”网络，把自己的能力教给一个“学生”网络——更小、更平行、更适用于现代计算机硬件的神经网络
- 比WaveNet网络的速度提升不少



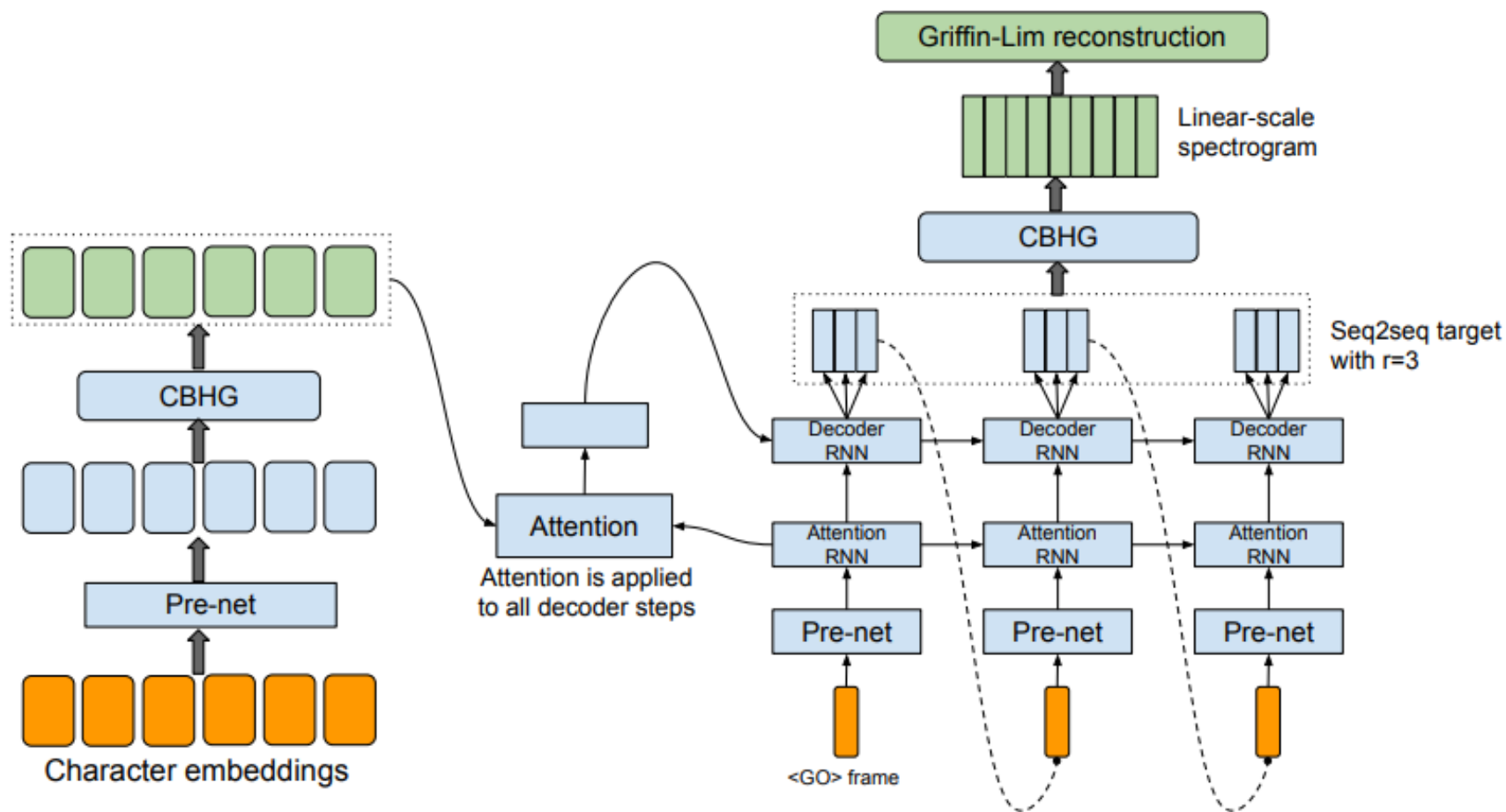
# 语音合成模型：Tacotron1.0

---

- ❑ TACOTRON是一个端到端的深度学习TTS模型
- ❑ 我们不必花费大量的时间去了解TTS中需要用的模块或者领域知识，直接用深度学习的方法训练出一个TTS模型
- ❑ 模型训练完成后，给定input，模型就能生成对应的音频



# 语音合成模型：Tacotron1.0





# 语音合成模型：Tacotron 2.0

---

- ❑ Tacotron2.0利用了谷歌此前在语音生成方面最强大的两种技术：WaveNet和Tacotron 1.0
- ❑ Tacotron 2使用文本和文字叙述来计算所有语言规则，而不再需要人工明确告知系统规则
- ❑ 文本本身被转换为Tacotron风格的“梅尔频谱”，实现节奏和强调。而单词本身则基于WaveNet风格的系统来生成



# 语音合成模型：ClariNet

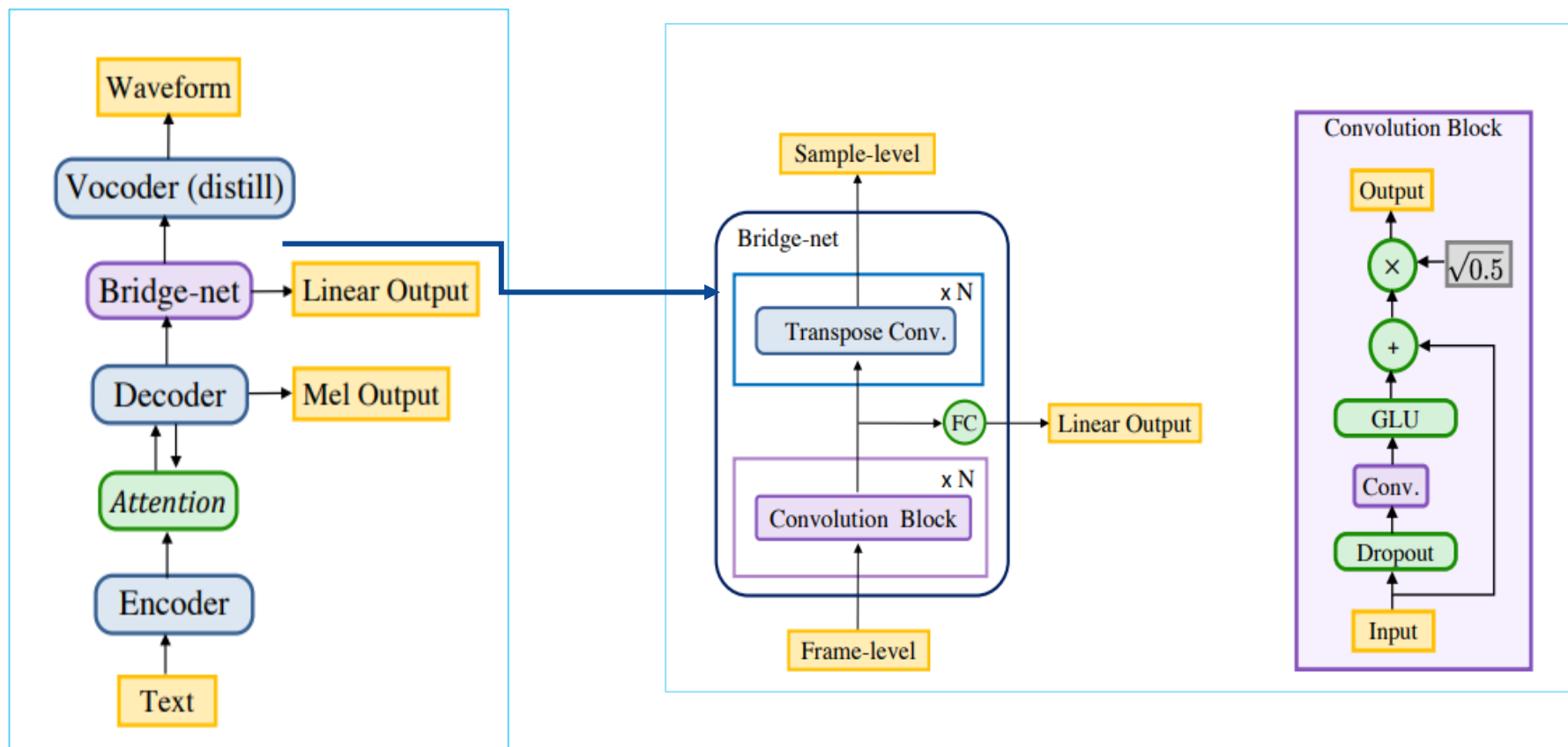
- ClariNet是由百度2018年提出的语音合成领域第一个完全端到端的系统



<https://clarinet-demo.github.io/>



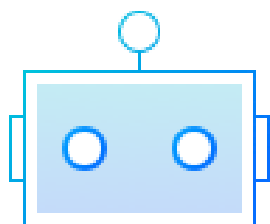
# 语音合成模型：ClariNet



# 语音合成的主要应用

## □ 语音交互

- 可集成到儿童故事机、智能机器人、平板设备等智能硬件设备，使用户与设备的交互更自然、更亲切



机器人回复



高拟真度合成



自然人机交互

# 语音合成的主要应用

## □ 有声阅读

- 通过阅读类APP阅读小说或新闻时，使用语音合成技术为用户提供多种发音人的朗读功能，释放双手和双眼，获得更极致的阅读体验



电子教材



合成朗读音频



丰富学习途径

# 语音合成的主要应用

## □ 语音播报

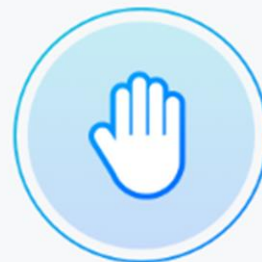
- 可应用于打车软件、餐饮叫号、排队软件等场景，通过语音合成进行订单播报，让您便捷获得通知信息



地图导航



语音播报



解放双手



# 6

## 中英文术语对照



# 中英文术语对照

---

- ❑ 语音识别: **Speech Recognition**
- ❑ 声学模型: **Acoustic Model**
- ❑ 语言模型: **Language Model**
- ❑ 分帧: **Frame Blocking**
- ❑ 预加重: **Pre-emphasis**
- ❑ 加窗: **Windowing**
- ❑ 线性预测系数: **Linear Prediction Coefficients, LPC**
- ❑ 线性预测倒谱系数: **Linear Prediction Cepstral Coefficients, LPCC**





# 中英文术语对照

---

- ❑ 梅尔频率倒谱系数: Mel Frequency Cepstral Coefficients, MFCC
- ❑ 高斯混合模型: Gaussian mixture model
- ❑ 隐马尔科夫模型: Hidden Markov model
- ❑ 快速傅里叶变换: Fast Fourier Transform, FFT
- ❑ 离散余弦变换: Discrete Cosine Transform, DCT
- ❑ 深度全序列卷积神经网络: Deep Fully Convolutional Neural Network
- ❑ 联结时序分类: Connectionist Temporal Classification, CTC



# 中英文术语对照

---

- 语音合成: Text To Speech/ Speech Synthesis
- 文本分析: Text Analysis
- 声纹识别: Voice Print Recognition
- 极大后验概率: Maximum A Posteriori
- 通用背景模型: Universal Background Model
- 高斯超向量: Gaussian Super Vector
- 身份认证向量: Identity-Vector



# 谢谢！

