北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

语音文字识别和说话人识别

小组成员：王虔翔 王凯 梁晨 和昕 钱鹏屿

2021.9.27

CONTENTS

1 语音文字识别经典综述

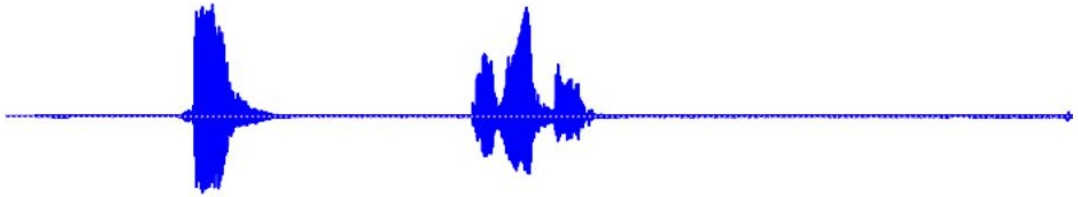
汇报人：王虔翔

1.1 发展历程

年代	发展	说明
1950	孤立词识别	1952年贝尔实验室首次实现Audrey英文数字识别系统，可以识别单个数字0～9的发音，并且对熟人的准确度高达90%以上。
1970	大词汇量的孤立词识别	卡耐基梅隆大学研发出harpy语音识别系统，该系统能够识别1011个单词
1980	发展到连续词识别	出现隐马尔科夫模型（**HMM**）、**N-gram**语言模型等现在依然使用的重要技术
1990	大词汇量连续词识别持续进步	提出了区分性的模型训练方法MCE和MMI和模型自适应方法MAP和MLLR
2010	深度学习模型	2006年，Hinton提出深度置信网络（DBN）；2009年， Hinton和学生Mohamed将深度神经网络应用于语音识别，在小词汇量连续语音识别任务TIMIT上获得成功
2020	端到端模型	**CTC**、**Seq2Seq**等模型持续发展

1.2 基本原理

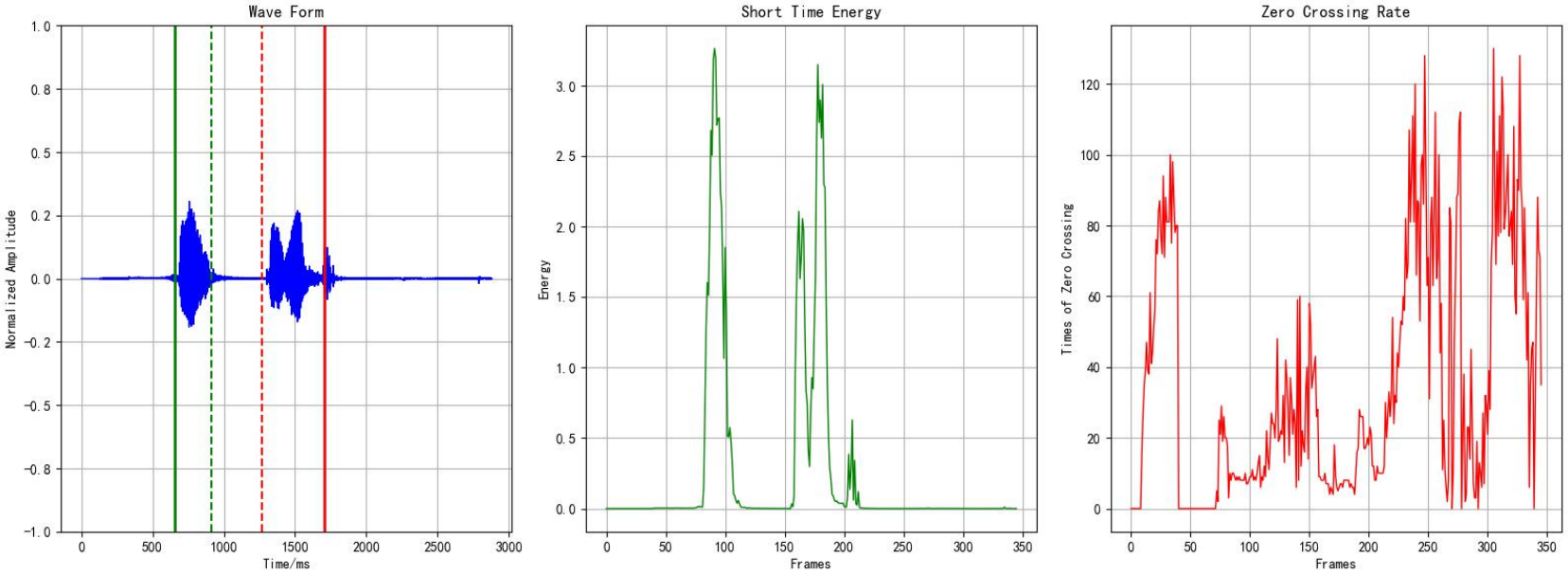
“嗨，大家好！”



预处理　　声学特征提取　　建立声学模型与语言模型

1.2 基本原理——预处理

(1) VAD (Voice Activity Detection)



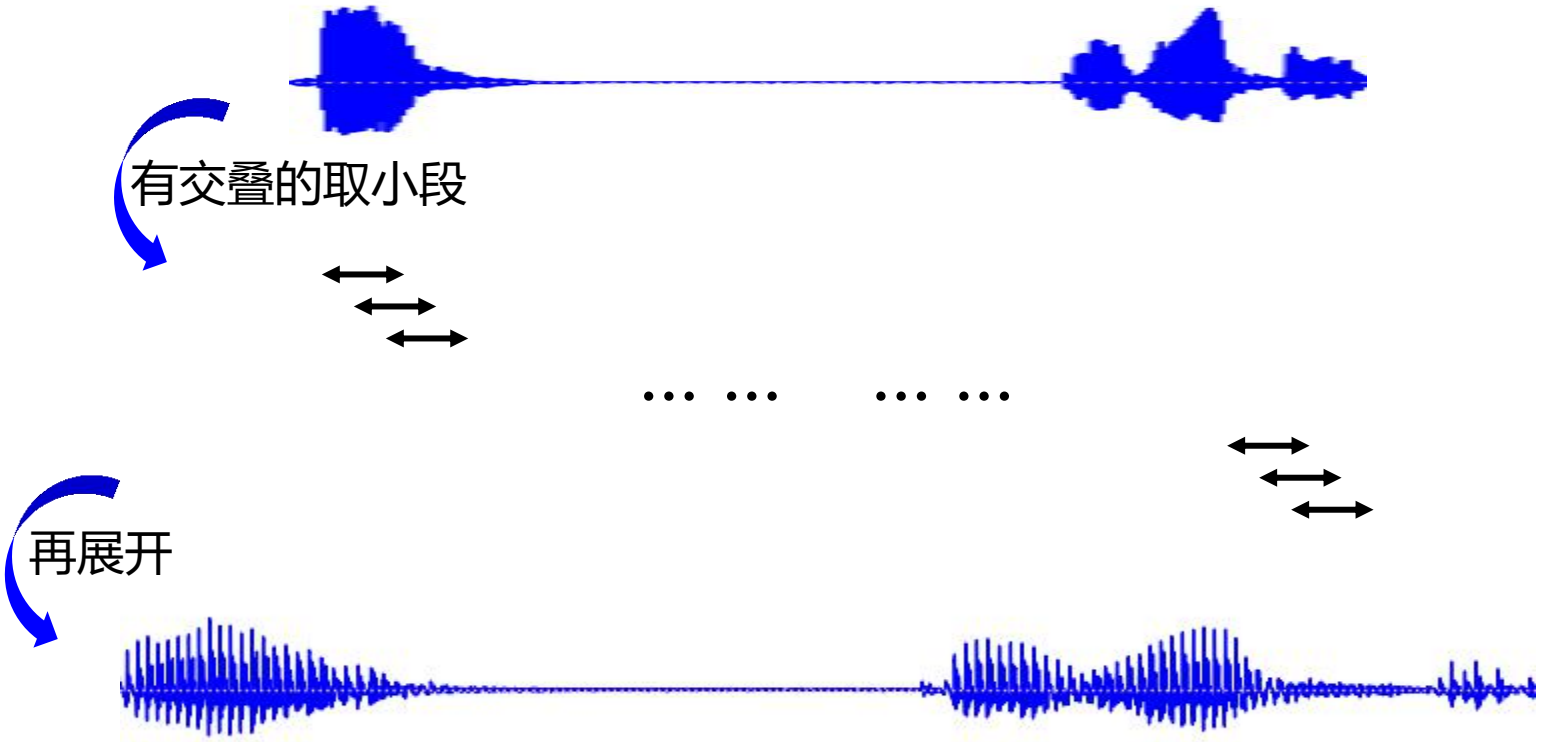
即，静音抑制或语音端点检测。

指的是从声音信号流里识别和消除长时间的静音期。

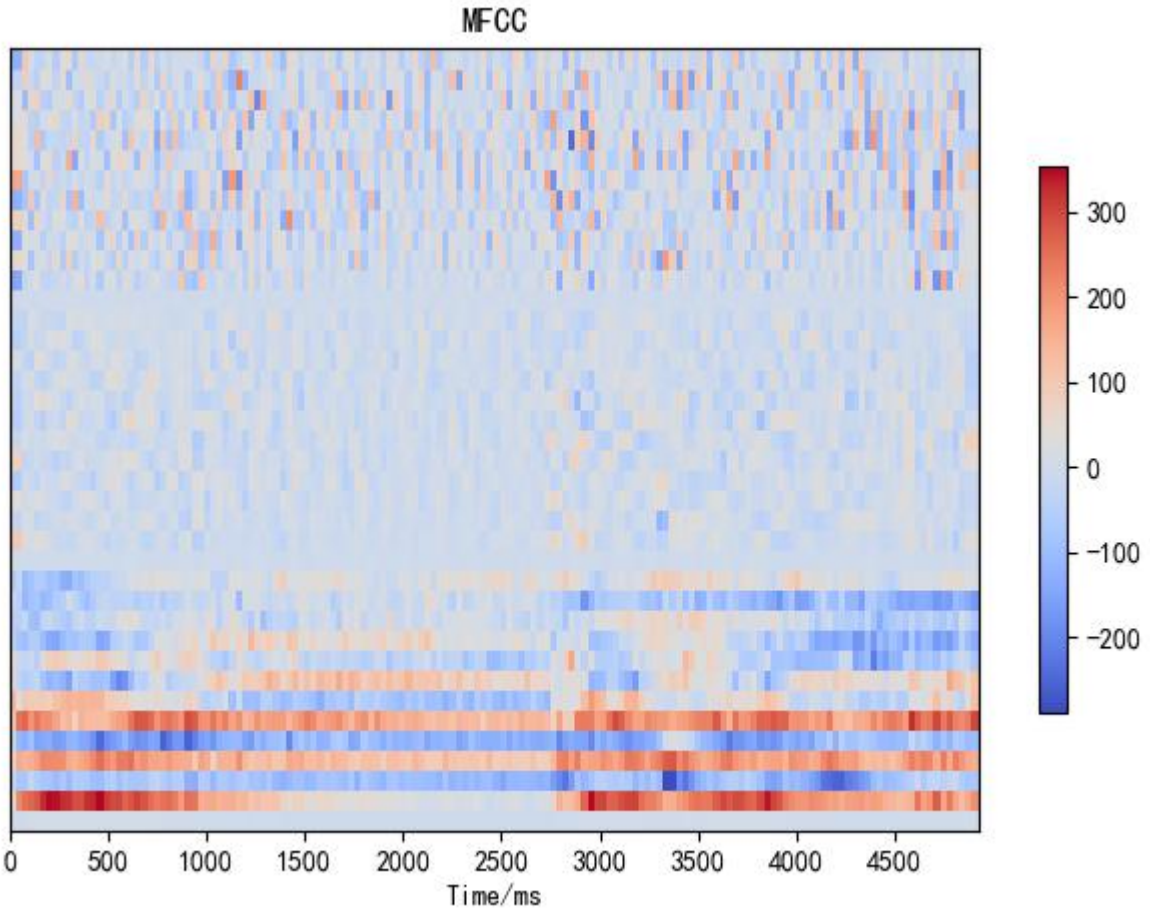
图中演示的是通过短时能量和过零率来切割下“嗨大家好”声音的首尾空白

1.2 基本原理——预处理

(2) 分帧



1.2 基本原理——声音特征提取



1.2 基本原理——建立声学模型与语言模型

文字序列　　语音输入

$$W^* = \arg\max_w P(W \mid Y) \quad (1)$$

$$= \arg\max_w \frac{P(Y \mid W)P(W)}{P(Y)} \quad (2)$$

$$\approx \arg\max_w P(Y \mid W)P(W) \quad (3)$$

Acoustic Model(AM)
声学模型

Language Model(LM)
语言模型

1.3 语言模型

语言模型通常用来衡量一个单词序列出现的概率。

“ hai	**嗨** 害 还 海 ”
haida	**还大** 还打 海大 海达
haidajia	**还打架** 害大家 还打假
haidajiahao	**嗨大家好** 还大家好 还打架好

T是由词序列$A_1,A_2,A_3,...A_n$组成的，即使得

$P(T)=P(A_1A_2A_3...A_n)$最大

1.3 语言模型

《季姬击鸡记》

季姬寂，集鸡，鸡即棘鸡。棘鸡饥叽，季姬及箕稷济鸡。鸡既济，跻姬笈，季姬忌，急咭鸡，鸡急，继圾几，季姬急，即籍箕击鸡，箕疾击几伎，伎即藟，鸡叽集几基，季姬急极屐击鸡，鸡既殛，季姬激，即记《季姬击鸡记》。

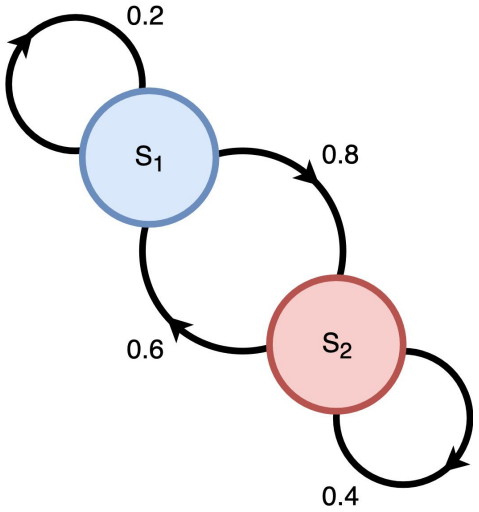
1.3 语言模型—— n-gram

“Haidajiahao | **嗨大家好** 还大家好 还打架好”

T是由词序列$A_1,A_2,A_3,...A_n$组成的，即使得

$P(T)=P(A_1A_2A_3...A_n)$最大

$P(T)=P(A_1A_2A_3...A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)...P(A_n|A_1A_2...A_{n-1})$

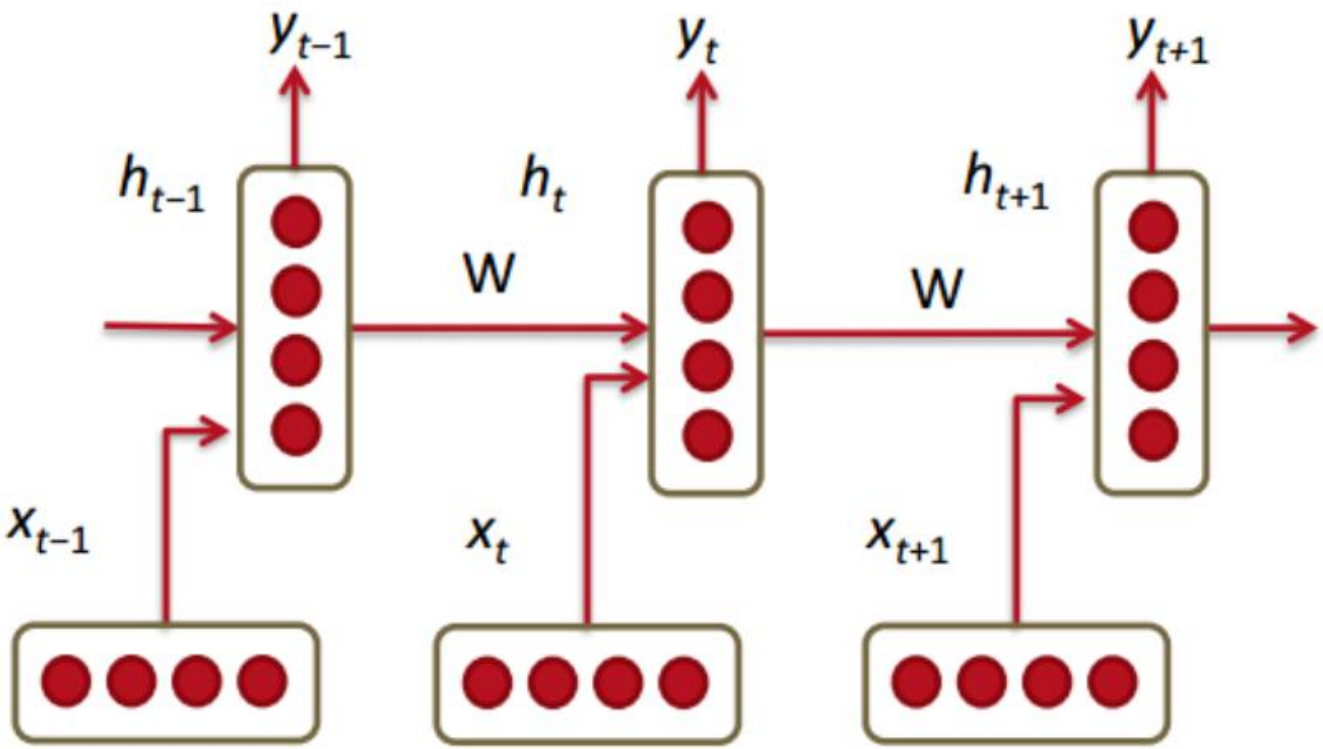


马尔科夫假设：一个item的出现概率，只与其前m个items有关

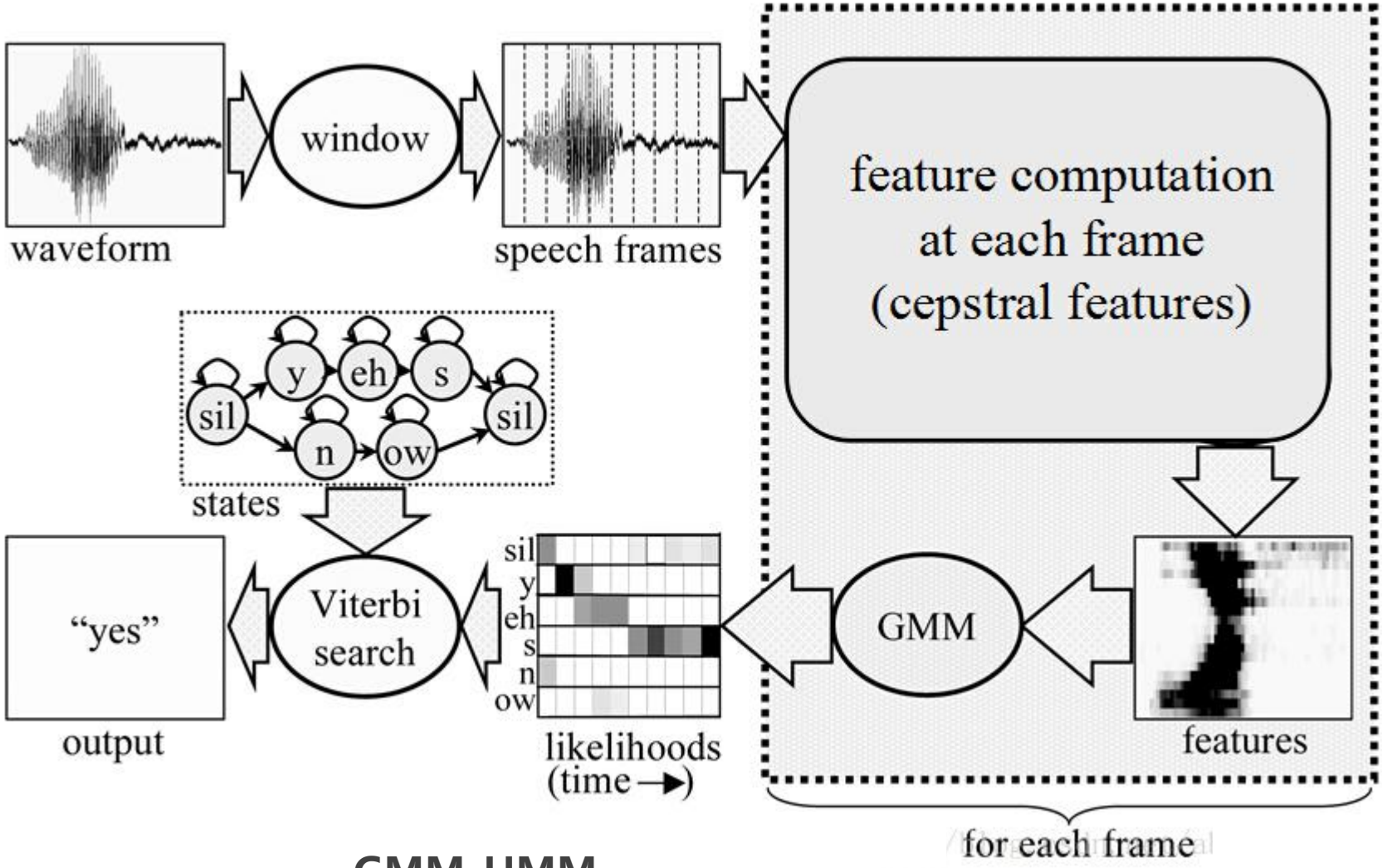
2-gram:

$P(T)=P(A_1)P(A_2|A_1)P(A_3|A_2)...P(A_n|A_{n-1})$

1.3 语言模型—— RNN



1.4 声学模型——传统模型

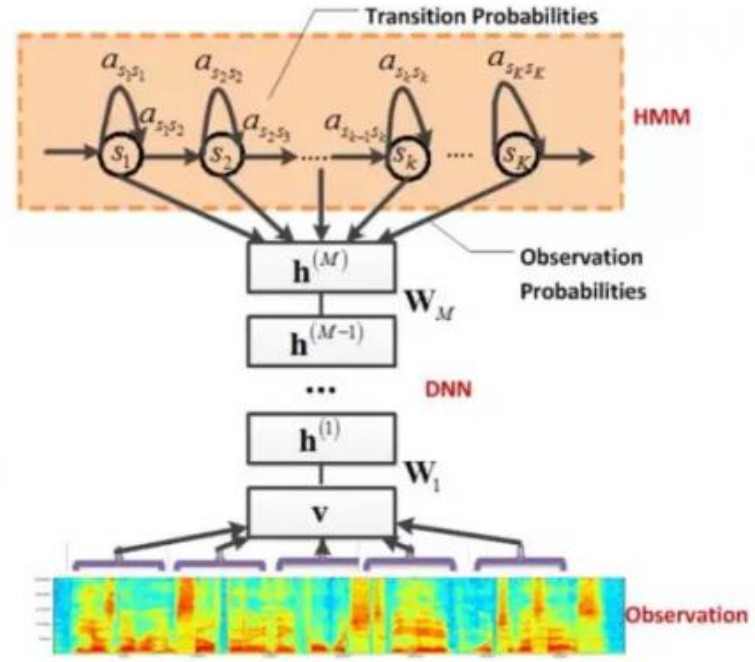


GMM-HMM

高斯混合模型和隐马尔科夫模型

1.4 声学模型——基于深度学习的模型

（1）DNN（深度神经网络）

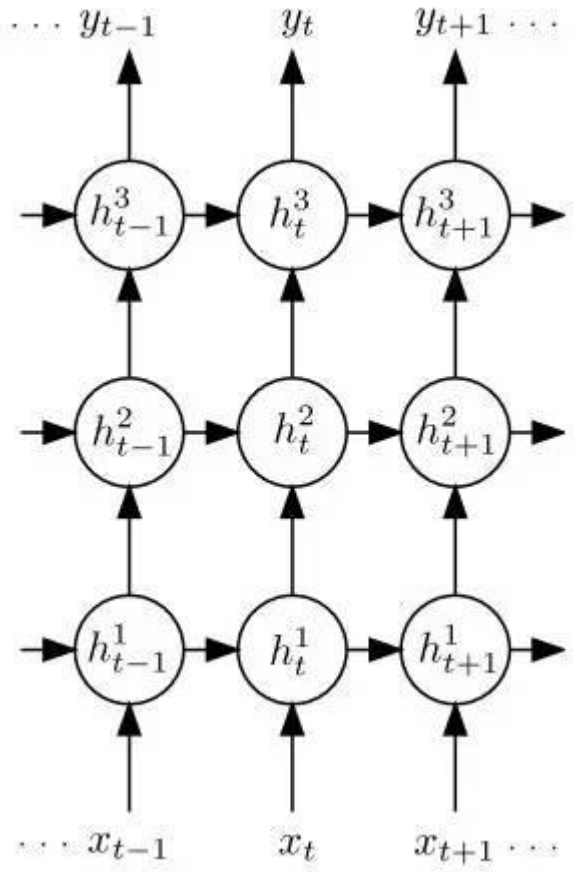


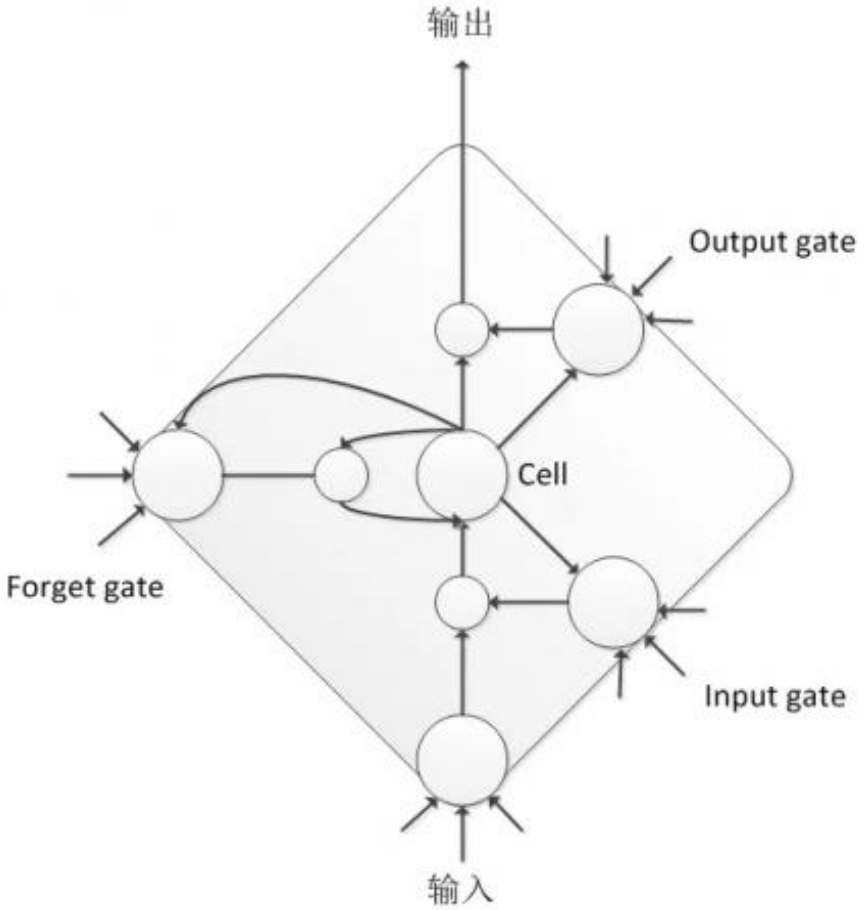
- 用DNN替换了GMM来对输入语音信号的观察概率进行建模
- 与GMM采用单帧特征作为输入不同，DNN将相邻的若干帧进行拼接来得到一个包含更多信息的输入向量
- 优点：

✓ DNN不需要对声学特征所服从的分布进行假设
✓ DNN由于使用拼接帧，可以更好地利用上下文的信息
✓ DNN的训练过程可以采用随机优化算法来实现，可以接受更大的数据规模

1.4 声学模型——基于深度学习的模型

(2) RNN与LSTM

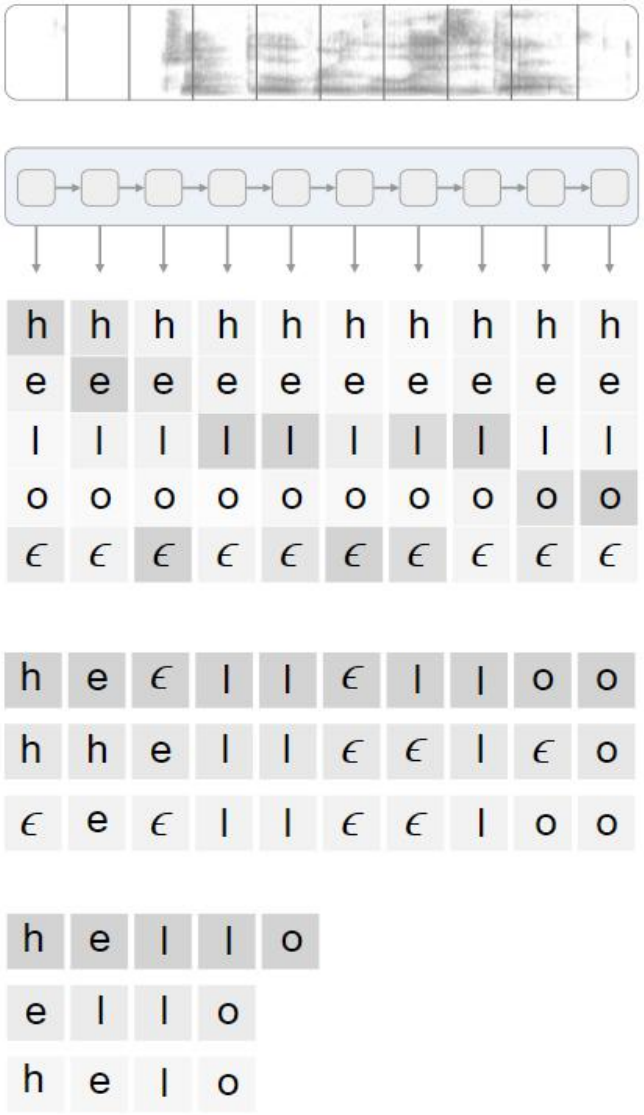




1.4 声学模型——端到端模型

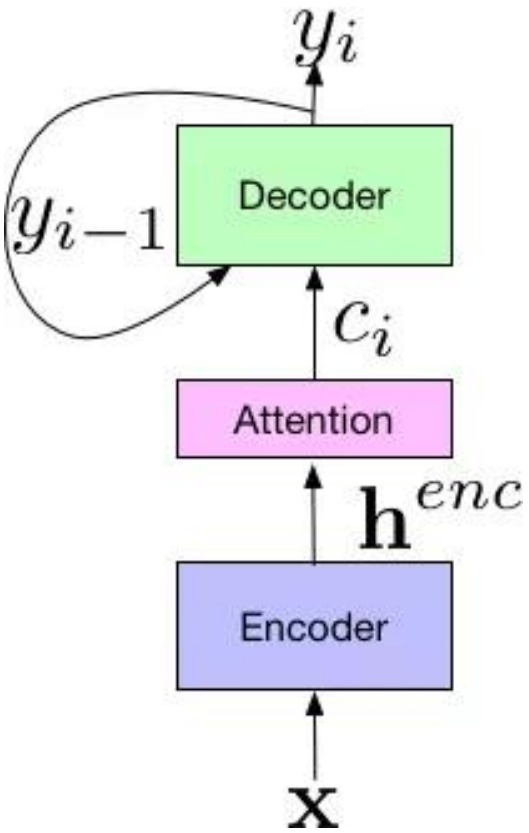
（1）CTC

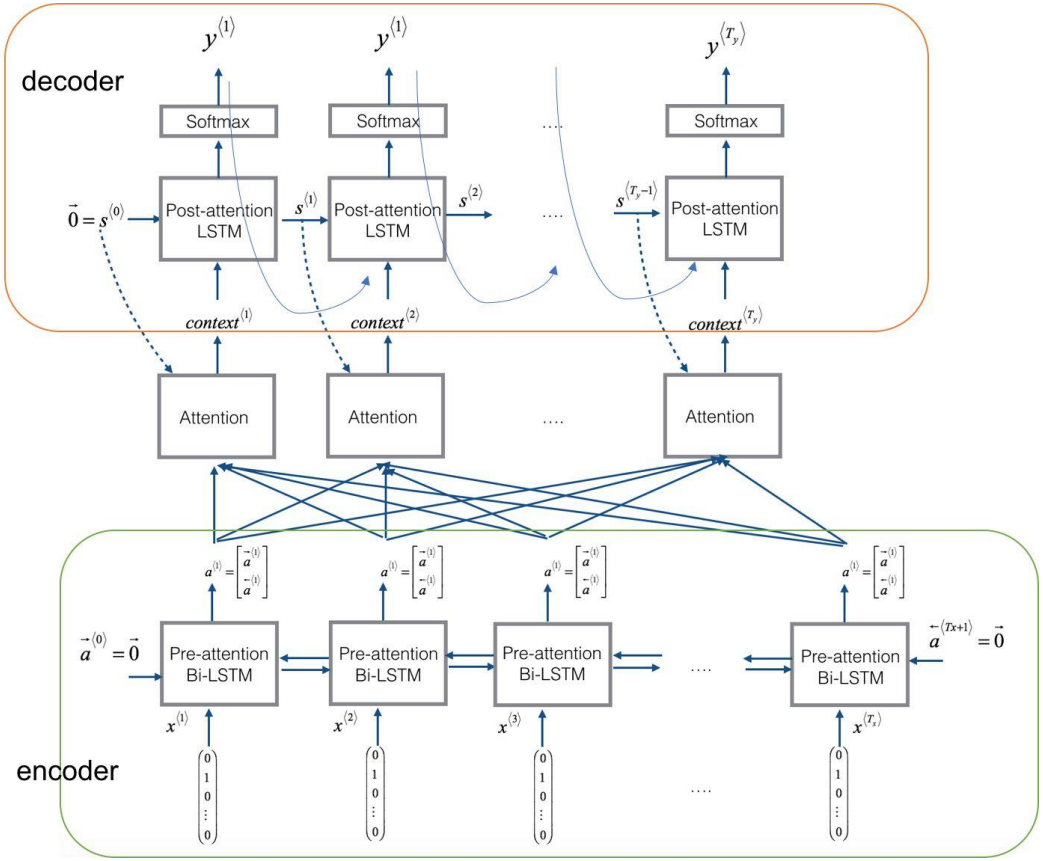
- Connectionist temporal classification 连接时序分类
- 主要用于处理序列标注问题中的输入与输出标签的对齐问题
- 引入了Blank
- 通过CTC，可以直接将语音在时间上的帧序列和相应的转录文字序列在模型训练过程中自动对齐



1.4 声学模型——端到端模型

（2） Sequence-to-Sequence





2 语音文字识别前沿进展

汇报人：王凯

2.1 语音识别技术的发展

2011年前后

基于DNN+HMM的语音识别。

2014年前后

基于LSTM+CTC的不完全端到端语音识别。

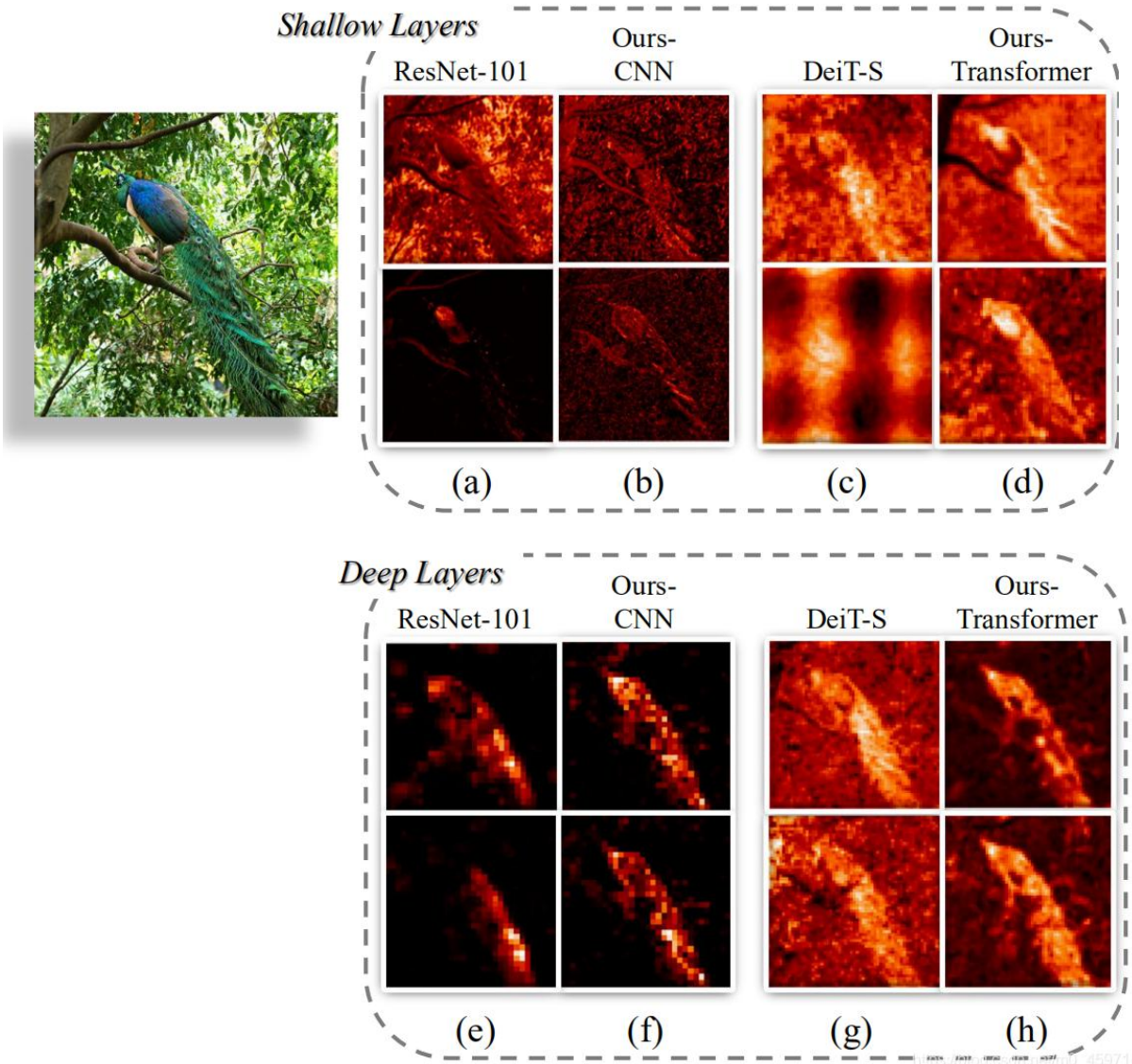
2017年前后

基于Transformer的完全端到端语音识别。

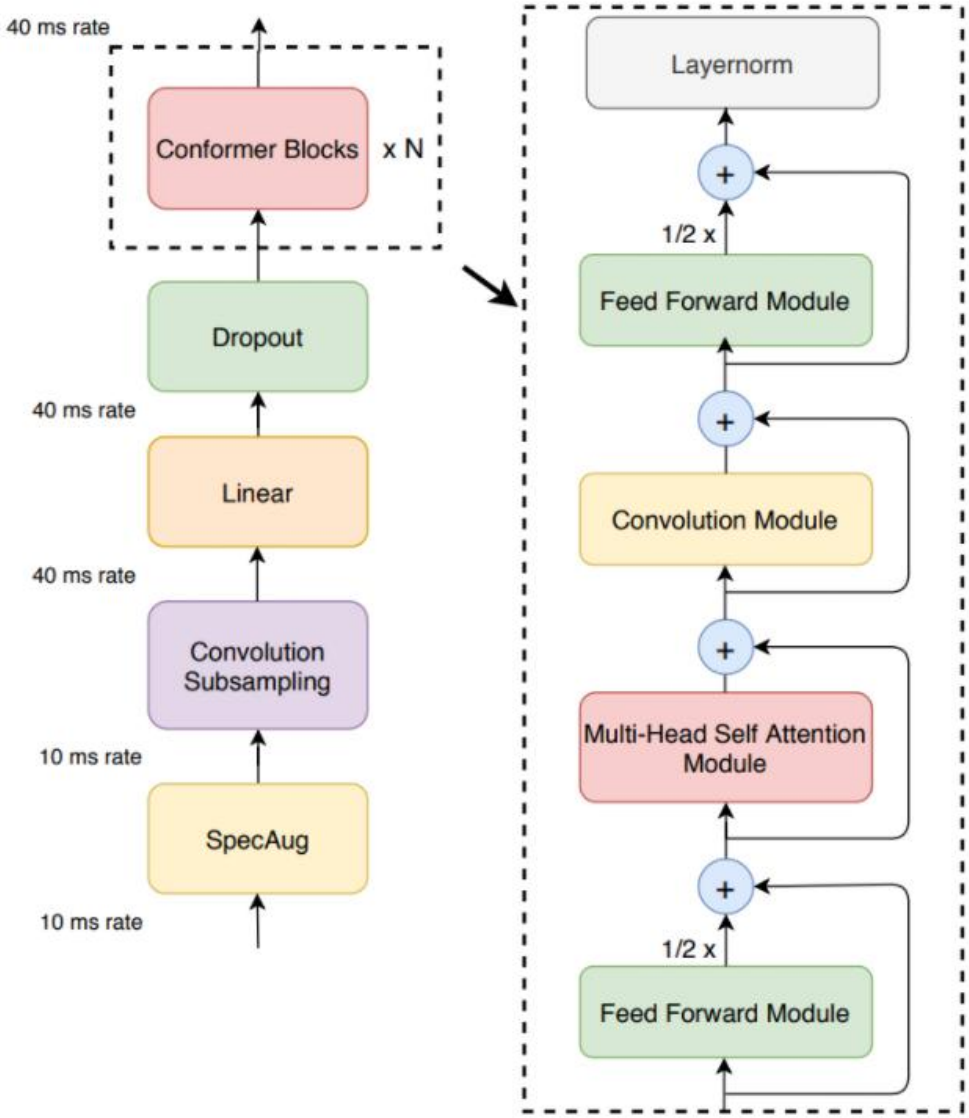
2020年

提出一种混合网络Conformer，充分利用到CNN和Transformer机制的优点。

2.2 Conformer——CNN增强的Transformer

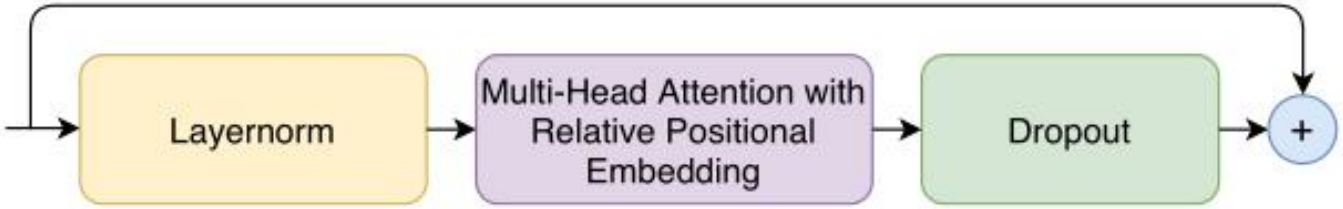


2.2 Conformer——CNN增强的Transformer

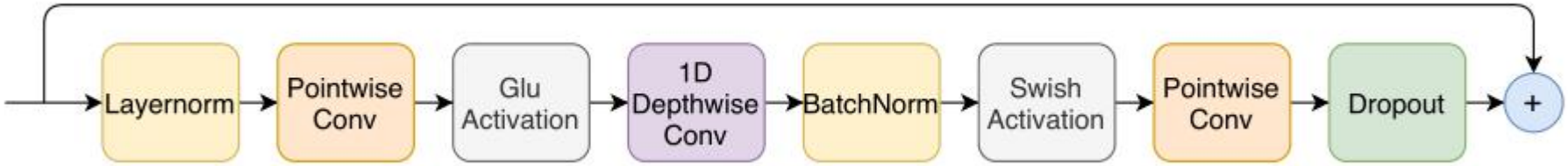


2.2 Conformer——CNN增强的Transformer

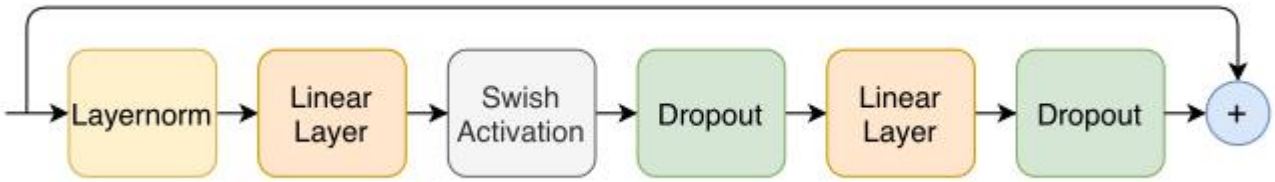
- Self-attention模块



- 卷积模块



- FF模块



2.2 Conformer——CNN增强的Transformer

- 三明治结构

$$\tilde{x}_i = x_i + \frac{1}{2}\mathrm{FFN}(x_i)$$
$$x'_i = \tilde{x}_i + \mathrm{MHSA}(\tilde{x}_i)$$
$$x''_i = x'_i + \mathrm{Conv}(x'_i)$$
$$y_i = \mathrm{Layernorm}(x''_i + \frac{1}{2}\mathrm{FFN}(x''_i))$$

2.2 Conformer——CNN增强的Transformer

Method	#Params (M)	WER Without LM		WER With LM	
		testclean	testother	testclean	testother
Hybrid					
Transformer [33]	-	-	-	2.26	4.85
CTC					
QuartzNet [9]	19	3.90	11.28	2.69	7.25
LAS					
Transformer [34]	270	2.89	6.98	2.33	5.17
Transformer [19]	-	2.2	5.6	2.6	5.7
LSTM	360	2.6	6.0	2.2	5.2
Transducer					
Transformer [7]	139	2.4	5.6	2.0	4.6
ContextNet(S) [10]	10.8	2.9	7.0	2.3	5.5
ContextNet(M) [10]	31.4	2.4	5.4	**2.0**	4.5
ContextNet(L) [10]	112.7	**2.1**	4.6	**1.9**	4.1
Conformer (Ours)					
Conformer(S)	10.3	**2.7**	**6.3**	**2.1**	**5.0**
Conformer(M)	30.7	**2.3**	**5.0**	**2.0**	**4.3**
Conformer(L)	118.8	**2.1**	**4.3**	**1.9**	**3.9**

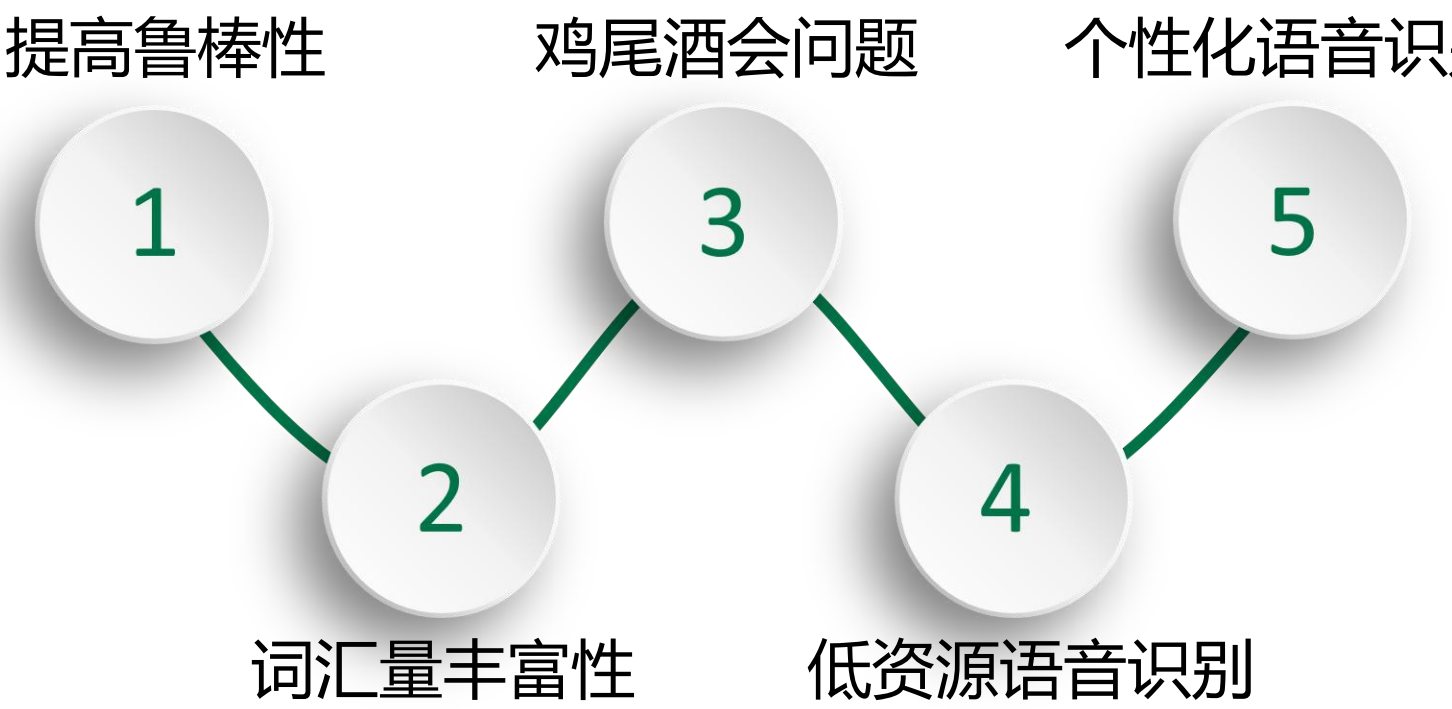
2.3 国内市场应用

公司简称	语音相关产品	相关技术	备注
科大讯飞	语音引擎（语音识别，语音合成，声纹识别） 教育产品（教学，考试，学习，智能硬件） 手机应用（讯飞输入法，灵犀，录音包，开心熊宝，听说无忧）；互动音乐等	语音识别，语音合成，语音唤醒，自然语言处理	国内四六级英语口语评测技术和录音质检软件提供商,识别准确率超过95%。支持中英粤藏维等5个语种、川豫和东北等方言。语音输入速度达180字/分，识别结果响应时间低于500ms，支持在线/离线命令词识别，准确率99%
百度语音	智能语音云平台	语音识别，语音合成，语音唤醒	离线在线融合，识别技术很强，识别准确率90%以上
腾讯（微信）	智能语音SDK	语音识别，语音合成，声纹识别	通用领域的语音识别率最高可以达到95%，实时率0.27；语音合成MOS值4.4；声纹识别准确率99%
阿里（阿里云）	智能语音交互	语音识别，语音合成，自然语言处理	识别准确率97%。
思必驰	智能车载，智能家居，智能机器人，语音输入法	语音识别，语音合成，自然语言处理	3米识别率94%以上，5米识别率92以上

2.3 国内市场应用

公司简称	语音相关产品	相关技术	备注
中科院声学所（中科信利）	语音识别，关键词检测，音频检索，语音合成，语种识别，音频水印，语音情绪识别	语音识别，关键词检测，音频检索，语音合成，语种识别，音频水印，语音情绪识别	对于朗读类型语音，识别准确率在90%以上，经过模型优化训练以后能达到95%
中科院自动化所（极限元）	智能教育，智能安全，泛娱乐，智能车载	语音识别，语音合成，自然语言处理	对于自然对话类型语音，识别准确率为80%，经过模型优化训练以后能够达到85%
云知声	智能电视方案，音乐搜索方案，视频搜索方案，购物搜索方案	语音识别，语义理解，语音合成，音频转写	支持大词汇量连续语音在线识别，识别准确率可达99%以上
捷通华声（灵云）	灵云智能语音分析平台	语音合成，语音识麦克风阵列，语义理解，机器翻译，声纹识别	识别准确率超过96%
出门问问	问问手表，问问魔镜，问问耳机，问问音箱 手表操作系统 虚拟个人助手	语音识别，语音合成，自然语言处理	识别正确率96%，支持实时识别
驰声	智能语音评测技术，微口语解决方案，英语学习行业解决方案，英语正式/教辅考试方案	语音识别，语音合成，语音评测，对话管理	拥有自主研发的全套语音技术，包括在线语音识别，离线语音识别和离线热词识别，识别率高达95%

2.4 今后的一些研究方向



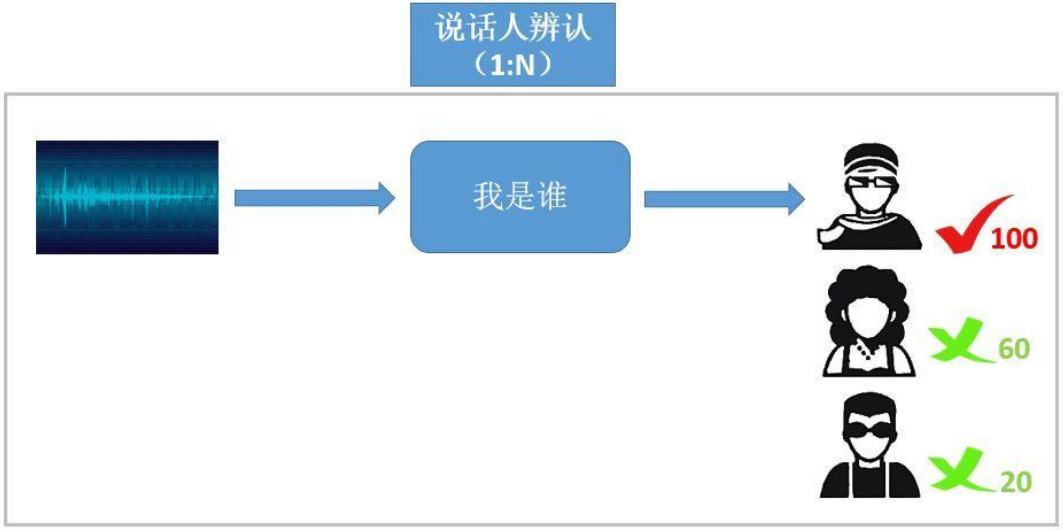
3 说话人识别经典综述

汇报人：梁晨

3.1 说话人识别-Speaker Identification (SI)

说话人识别是一个流行的研究课题，有各种应用，如安全、取证、生物特征认证。目前在这个领域有很多的研究，很多的方法被提出来，所以这个领域的最先进技术已相当成熟。

从一组已知的说话者中识别一个未知的说话者的任务：找到声音最接近测试样本的说话者。



3.2 发展进程

- 1963年，Bell 实验室的 S. Pruzansky 提出了基于**模板匹配**(template matching)和统计方差分析的说话人识别方法
- 从 20 世纪 70 年代末至 80 年代末，LPC、LPCC、DTW 动态时间规整法、VQ 矢量量化法、HMM 隐马尔科夫模型

模板匹配

- 20 世纪 90 年代以后，高斯混合模型GMM、GMM-UBM、GMM-SVM 、JFA 、i-vector 模型

统计模型

- 在2014年，...

3.3 传统统计模型-高斯混合模型GMM

高斯混合模型是具有如下形式的概率分布模型：

- GMM公式

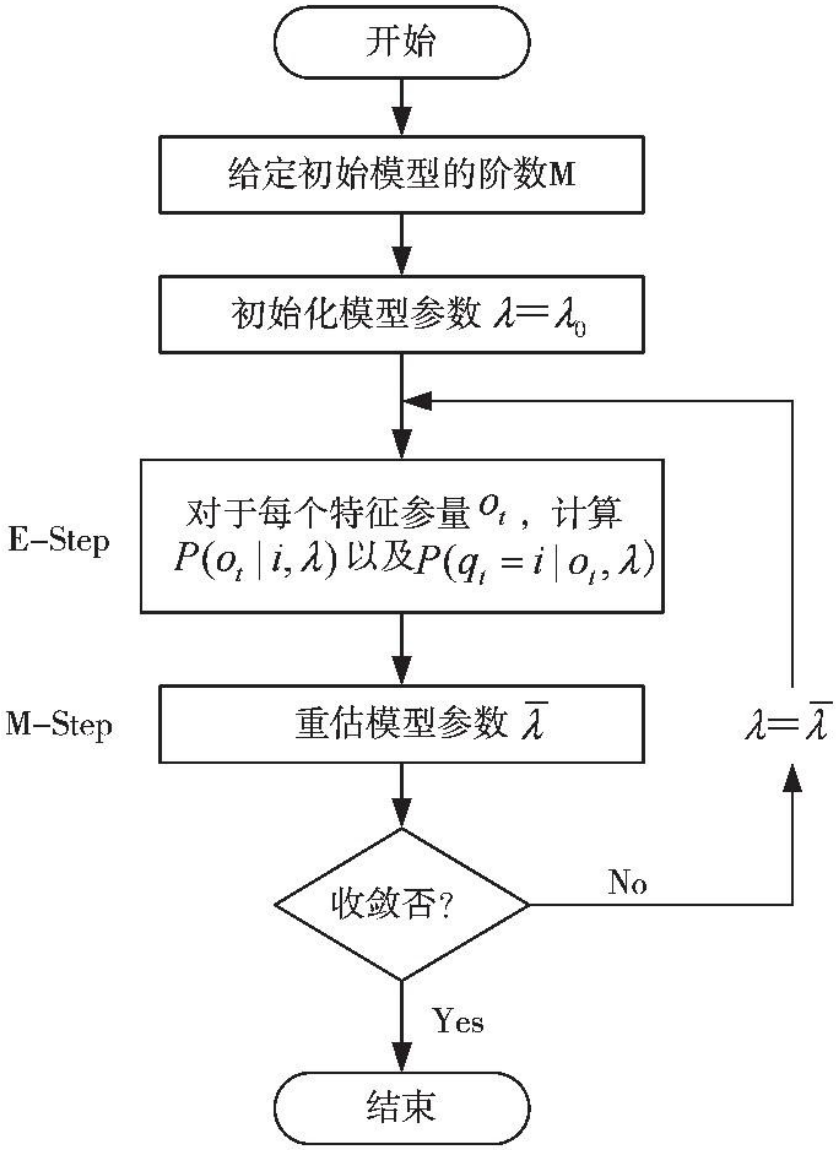
$$P(\mathrm{y}|\theta) = \sum_{k=1}^{K} \alpha_k \varphi(y|\theta_k)$$

- 高斯分布公式

$$\varphi(y|\theta_k) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

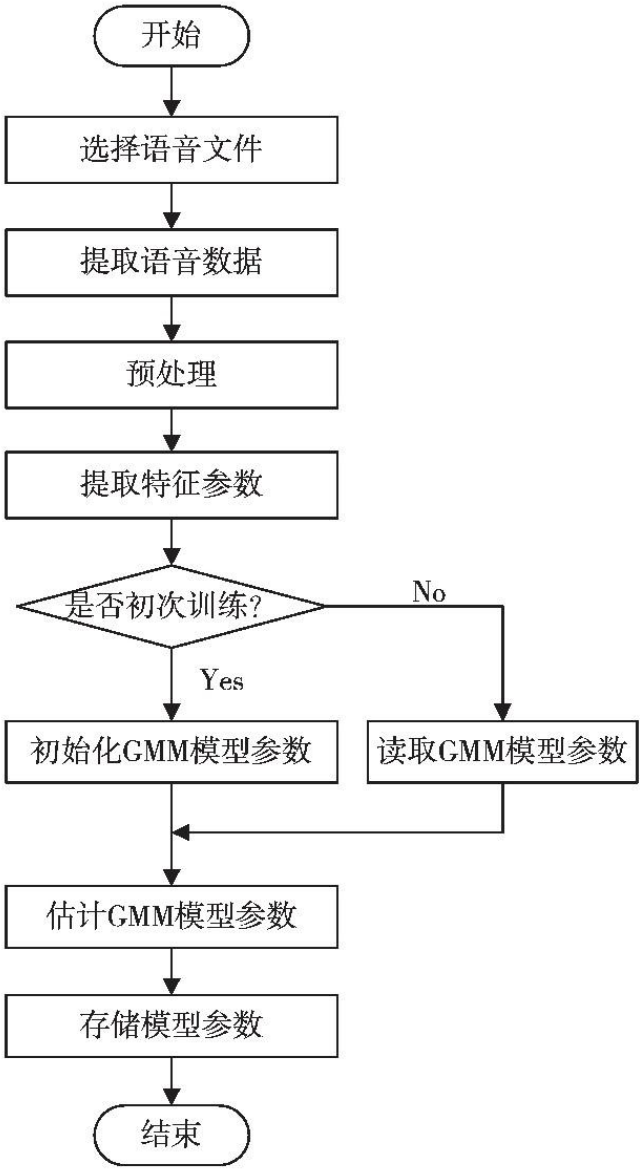
3.3 传统统计模型-高斯混合模型GMM

EM算法估计
GMM模型参数流程



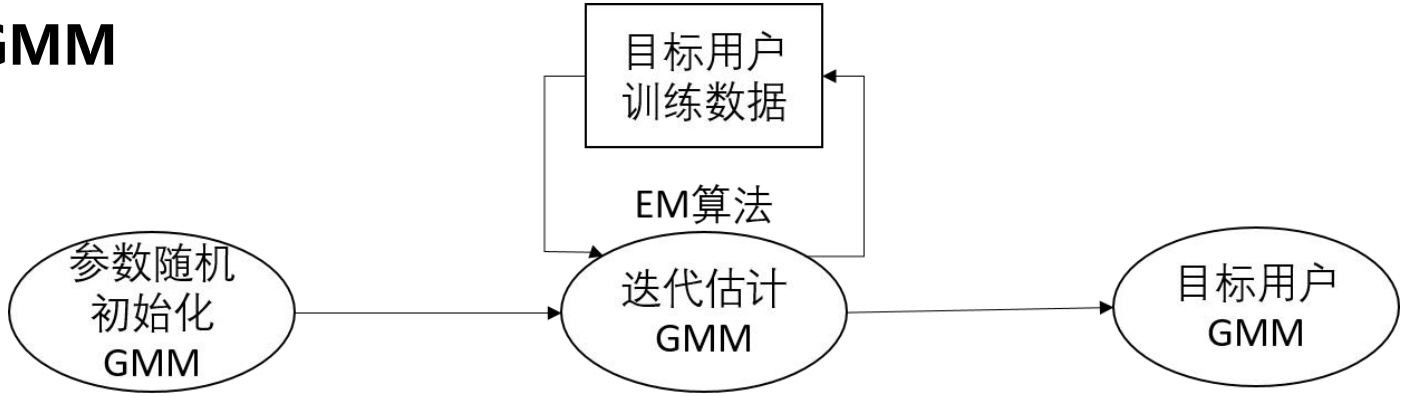
3.3 传统统计模型-高斯混合模型GMM

GMM模型训练流程

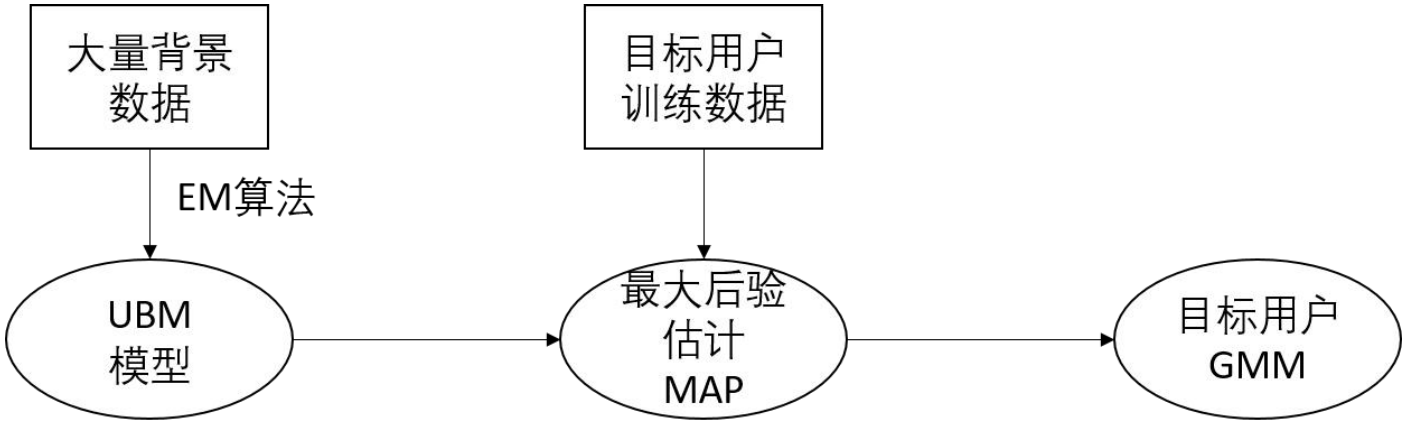


3.3 传统统计模型-GMM-UBM

- **GMM**



- **GMM-UBM**



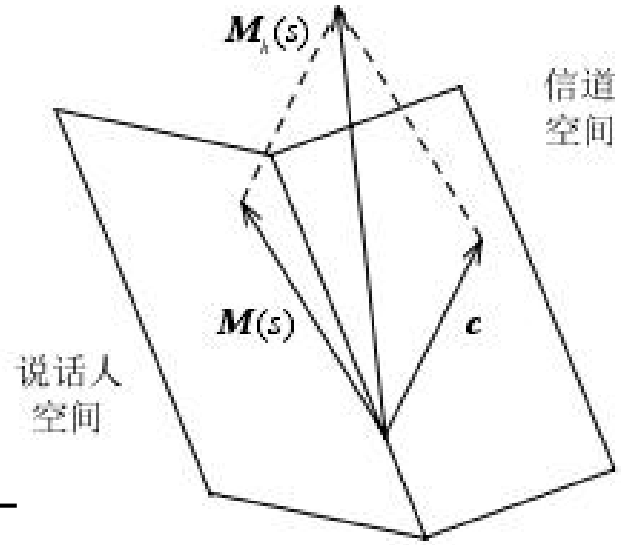
3.3 传统统计模型-JFA

说话人超矢量

$$\mathrm{M}(s) = \mathrm{m}_{ubm} + \mathrm{vy(s)} + \mathrm{dz(s)}$$

信道超矢量

$$\mathrm{c} = \mathrm{u}x_h(s)$$



说话人超矢量的组成

$$M_h(s) = \mathrm{m}_{ubm} + \mathrm{vy(s)} + \mathrm{dz(s)} + \mathrm{u}x_h(s)$$

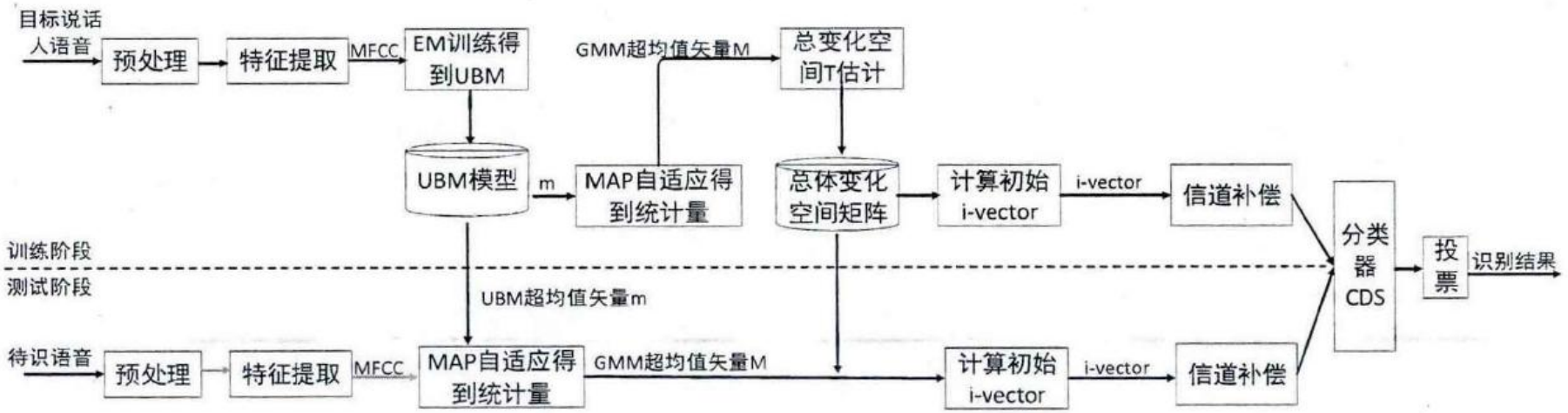
v: 低维说话人空间　　d: 残差矩阵

y(s) : 说话人因子　　z: 残差因子

u: 低维的信道空间　　m: UBM均值超矢量

$x_h(s)$: 第h段话所具有的信道因子

3.3 传统统计模型-i-vector



$$M_j(s) = m + Tw_j$$
$$w_j \sim N(0,1)$$

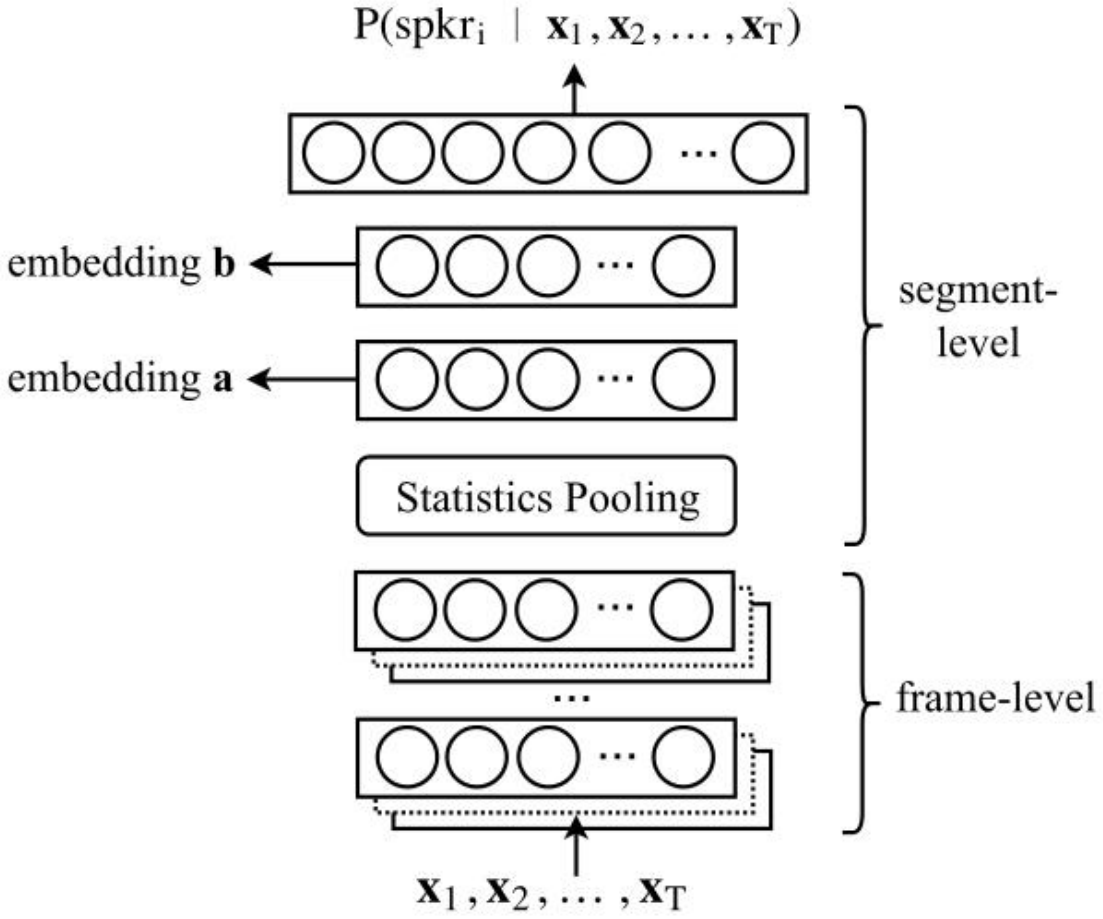
- M_j是标识某个训练说话人的第j句话的GMM均值超向量
- m用来表示UBM均值超向量
- w_j是标识总体变化子空间因子

4 说话人识别前沿进展

汇报人：和昕

4.1 前沿技术-x-vector

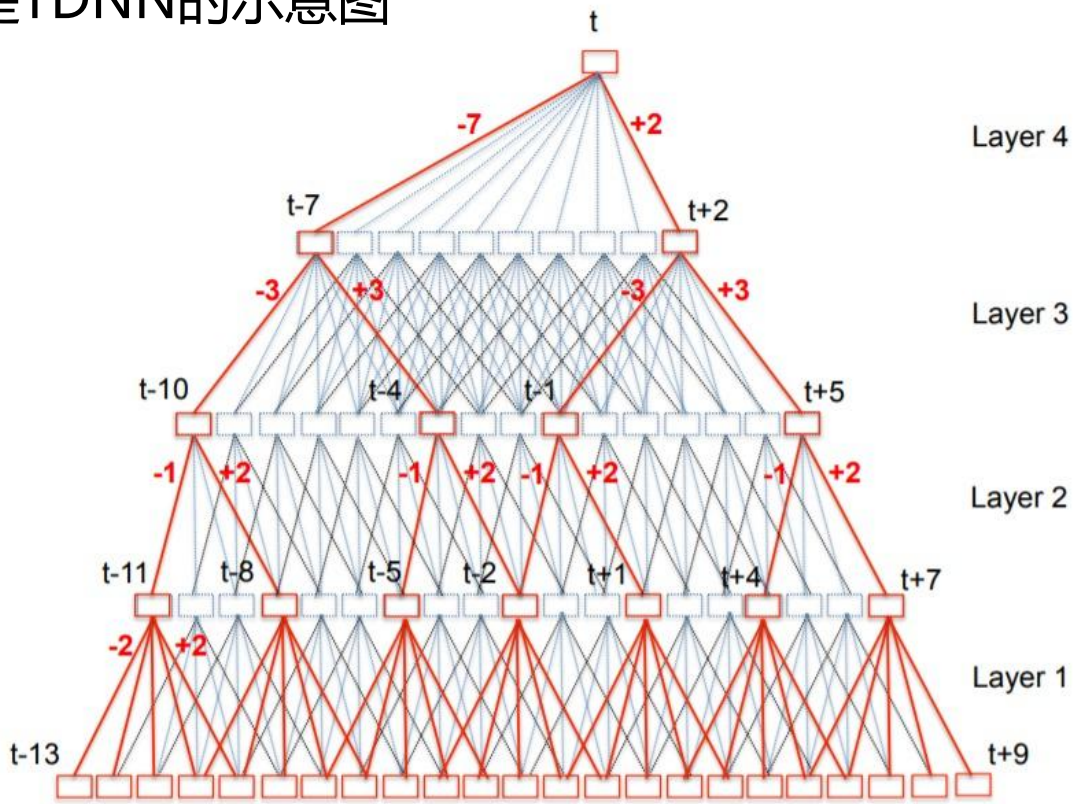
- 基本框架



4.1 前沿技术-x-vector

（1）Time Delay Neural Networks (TDNN)

系统框架图中的Statistics Pooling之前的部分就是TDNN，下图是TDNN的示意图



4.1 前沿技术-x-vector

(2) Statistics pooling

按照系统框架流程，pooling之后，使用两层全连接层，最后softmax之后的输出长度是说话人的数量K，模型的损失函数为下式的多分类交叉熵：

$$E = -\sum_{n=1}^{N}\sum_{k=1}^{K} d_{nk}\ln(P(spkr_k|\boldsymbol{x}_{1:T}^{(n)}))$$

- 其中，有K个说话人，每人说了N个片段(segment)
- 当对说话人片段n打上的标签是k时，d_{nk}是1，否则为0
- $P(spkr_k|\boldsymbol{x}_{1:T}^{(n)})$是输入语音n后，由softmax层给出的它属于各个说话人的后验概率

4.1 前沿技术-x-vector

（3） extract embedding

下图是i-vector、两种embedding方法以及联合两种embedding的对比结果：

- 测试语音比较长时，i-vector的优势比较明显；
- 测试语音在5-20s时，DNN效果较好；
- embedding的综合要比只用单个embedding好；
- DNN对out of domain 的效果优于i-vector

Table 1: *EER(%) on NIST SRE10*

	10s-10s	5s	10s	20s	60s	full
ivector	11.0	9.1	6.0	3.9	2.3	1.9
embedding **a**	11.0	9.5	5.7	3.9	3.0	2.6
embedding **b**	9.2	8.8	6.6	5.5	4.4	3.9
embeddings	7.9	7.6	5.0	3.8	2.9	2.6
fusion	8.1	6.8	4.3	2.9	2.1	1.8

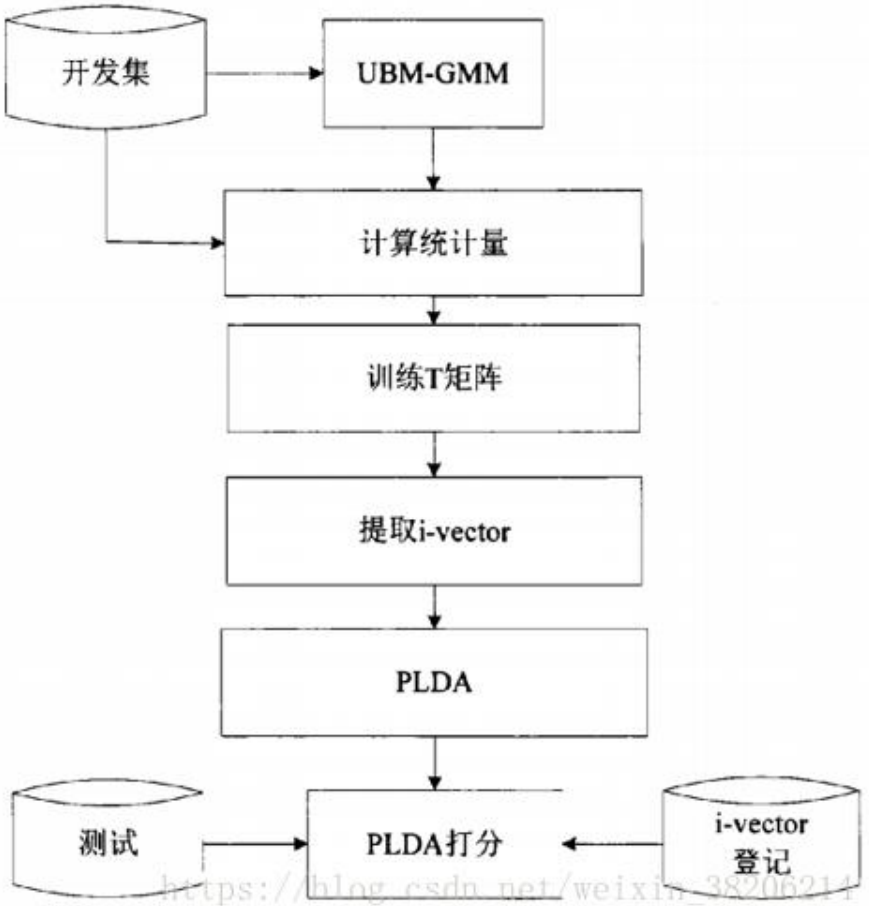
Table 3: *EER(%) on NIST SRE16*

	Cantonese	Tagalog	pool
ivector	8.3	17.6	13.6
embedding **a**	7.7	17.6	13.1
embedding **b**	7.8	17.4	13.1
embeddings	6.5	16.3	11.9
fusion	6.3	15.4	11.3

4.1 前沿技术-x-vector

(4) PLDA backend

PLDA(Probabilistic Linear Discriminant Analysis)是一种信道补偿算法，号称概率形式的LDA算法，PLDA算法的信道补偿能力比LDA更好，已经成为目前最好的信道补偿算法。



4.1 前沿技术-x-vector

(4) PLDA backend——建模

定义训练数据 x_{ij} 为第 i 个人的第 j 张图片(特征)，然后PLDA定义 x_{ij} 是按照下式生成的:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \epsilon_{ij}$$

由于我们只关心区分说话人，所以并不需要计算误差空间，所以可以把噪声部分的两项合并，则得到

$$X_{ij} = \mu + Fh_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \Sigma), h_i \sim N(0,1)$$

4.1 前沿技术-x-vector

（4） PLDA backend——训练

训练PLDA模型，就是通过训练数据估计出参数ϕ和Σ。估计这两个参数的方法是经典EM算法迭代求解，即猜（E-step)-反思（M-step），重复。

$$\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_N\end{bmatrix}=\begin{bmatrix}\mu\\ \mu\\ \vdots\\ \mu\end{bmatrix}+\begin{bmatrix}\mathbf{F} & \mathbf{G} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{F} & \mathbf{0} & \mathbf{G} & \dots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{h}\\ \mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \epsilon_N\end{bmatrix}$$

$$\mathbf{x}'=\mu'+\mathbf{A}\mathbf{y}+\epsilon'$$

训练流程：

1. 均值处理：计算所有训练数据的均值 μ，然后从训练数据中减去该均值
2. 初始化：初始化特征维度D，身份空间维度N_F，噪声空间维度N_G
3. EM迭代优化：训练过程中，先将上述模型写成矩阵形式

4.1 前沿技术-x-vector

(4) PLDA backend——训练

E-Step: 计算隐含变量 h 的期望

$$E[\mathbf{y}_i] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\Sigma'^{-1}(\mathbf{x}_i - \mu')$$
$$E[\mathbf{y}_i\mathbf{y}_i^T] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A}^T + \mathbf{I})^{-1} + E[\mathbf{y}_i]E[\mathbf{y}_i]^T$$

M-Step: 更新参数 $\theta = [\mu, F, G, \Sigma]$
更新的时候按照下述公式：

$$\mu = \frac{1}{IJ}\sum_{i,j}\mathbf{x}_{ij}$$

$$\mathbf{B} = \left(\sum_{i,j}(\mathbf{x}_{ij} - \mu)E[\mathbf{z}_i]^T\right)\left(\sum_{i,j}E[\mathbf{z}_i\mathbf{z}_i^T]\right)^{-1}$$

$$\Sigma = \frac{1}{IJ}\sum_{i,j}\mathbf{Diag}\left[(\mathbf{x}_{ij} - \mu)(\mathbf{x}_{ij} - \mu)^T - \mathbf{B}E[\mathbf{z}_i](\mathbf{x}_{ij} - \mu)^T\right]$$

4.1 前沿技术-x-vector

(4) PLDA backend——测试

模型测试阶段的思想是使用对数似然比来计算得分：

$$score = log\frac{p(n_1, n_2|H_s)}{p(n_1|H_d)p(n_2|H_d)}$$

得分越高，两条语音属于同一说话人的可能性越大

此外，在不使用PLDA或使用LDA信道补偿的情况下，使用余弦评分方法，来计算两个I-Vector矢量的相似度，以此来评分，具体计算公式如下：

$$score = \frac{< W_{tar}, W_{test} >}{\|W_{tar}\|\|W_{test}\|}$$

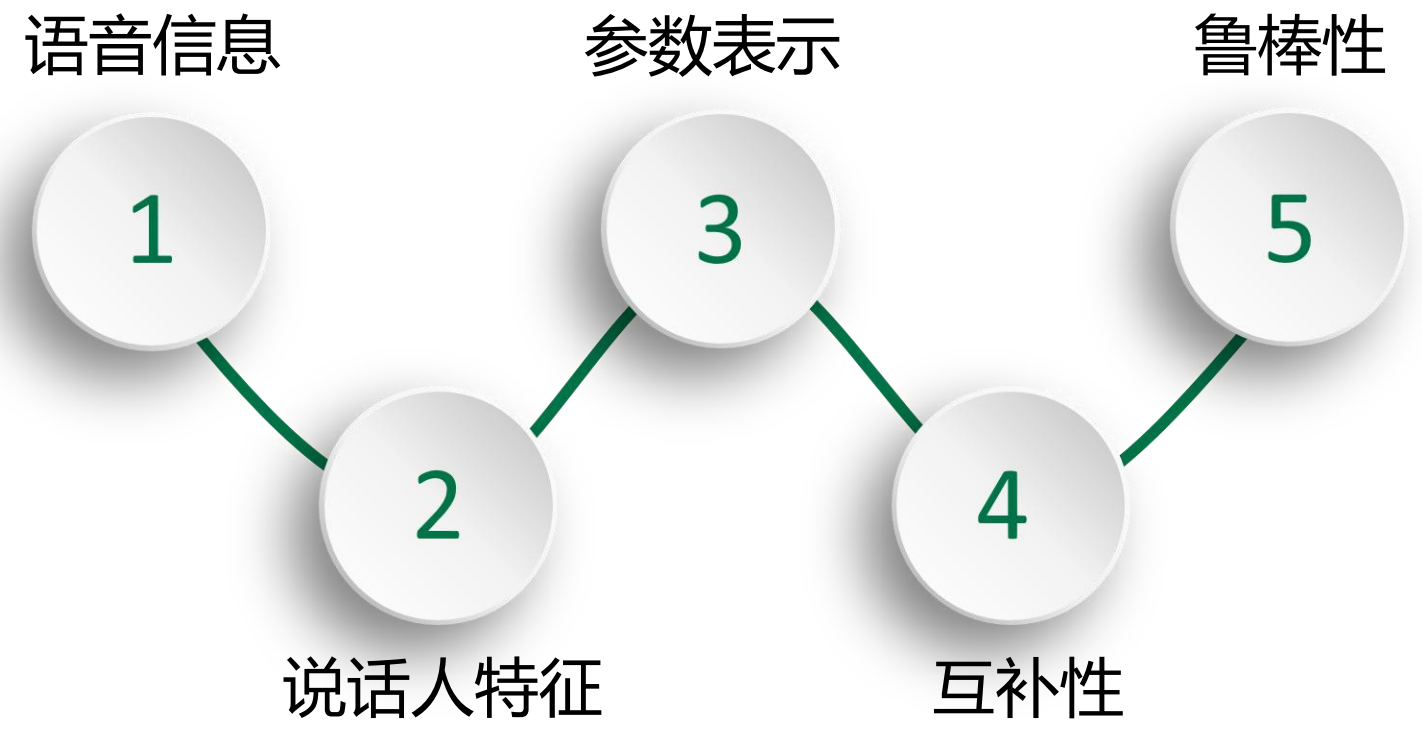
夹角反映了二者的相关性，当二者相关性大时，夹角小，分数高；当二者相关性小时，夹角大，分数低。

4.2 x-vector的相关应用

Kaldi的TIMIT全称The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus，是由德州仪器(TI)、麻省理工学院(MIT)和坦福研究院(SRI)合作基于x-vector技术构建的声学-音素连续语音语料库。

TIMIT语音库有着准确的音素标注，是一个学习用的好例子。在kaldi里面可以找到其语音识别的范例。

4.3 研究方向



5 demo展示

汇报人：钱鹏屿

5.1 demo——说话人识别

（1）数据简介

数据来自开源中文语音数据集AISHELL-1，总共178小时，语音来自400个讲话人，其中训练集340个人，测试解20个人，验证集40个人

在说话人识别demo中，选用AISHELL中的100人进行训练，并将数据分为5个fold做交叉验证

http://www.aishelltech.com/kysjcp

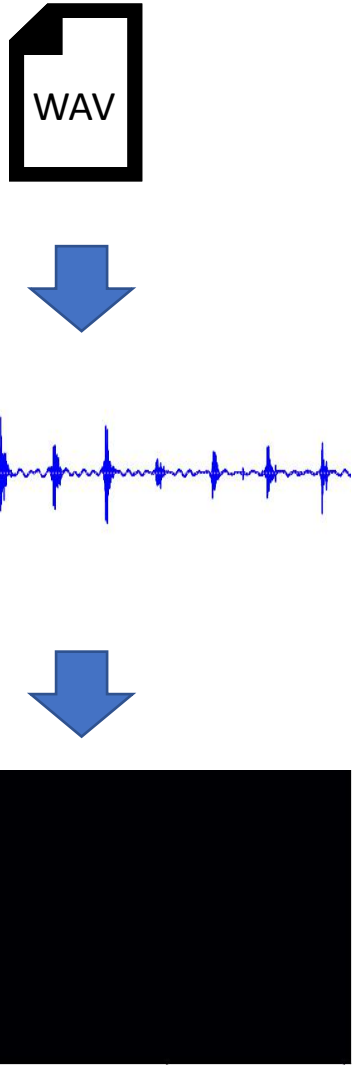
5.1 demo——说话人识别

（2）数据处理和音频特征提取

- librosa（音频信号处理）

librosa.load(file, sample_rate)
读入音频文件(.WAV)，采样率为sample_rate

librosa.feature.melspectrogram()
将音频时间序列转换为梅尔频谱特征(height*width)



梅尔频谱

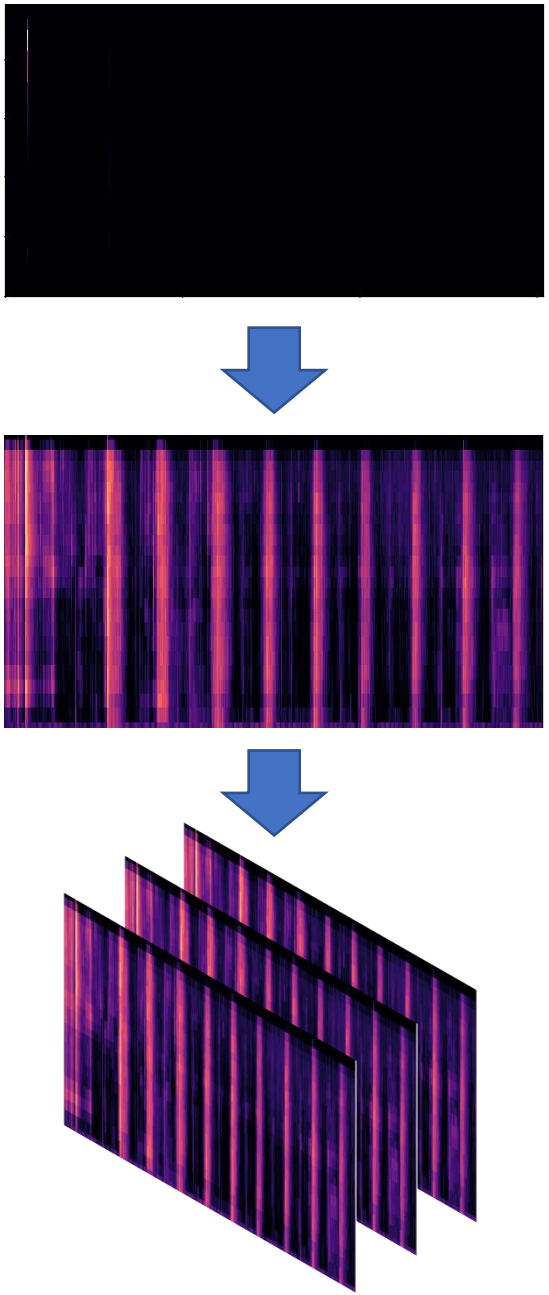
5.1 demo——说话人识别

（2）数据处理和音频特征提取

- librosa（音频信号处理）

librosa.power_to_db()
对梅尔频谱特征取对数

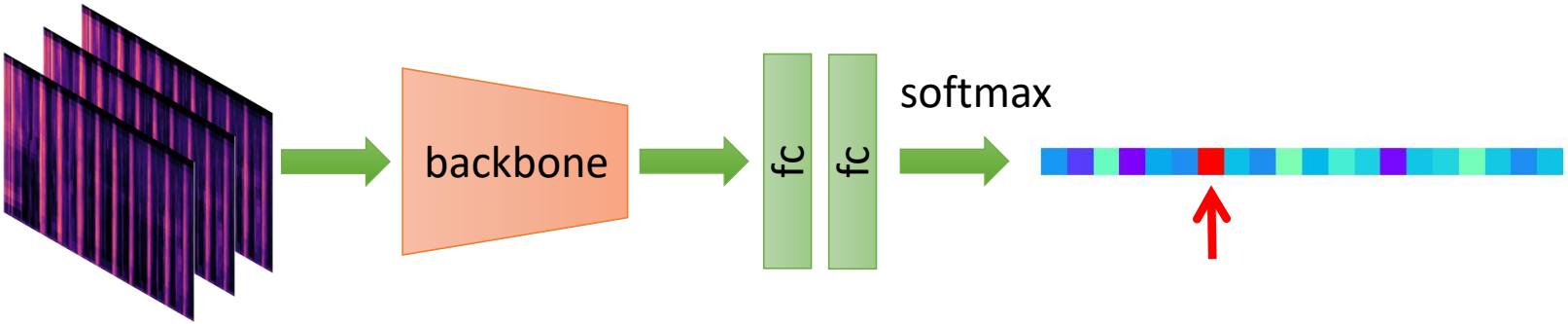
频谱图叠加三层形成RGB图像



5.1 demo——说话人识别

（3）分类器及分类效果

使用ResNet作为backbone。



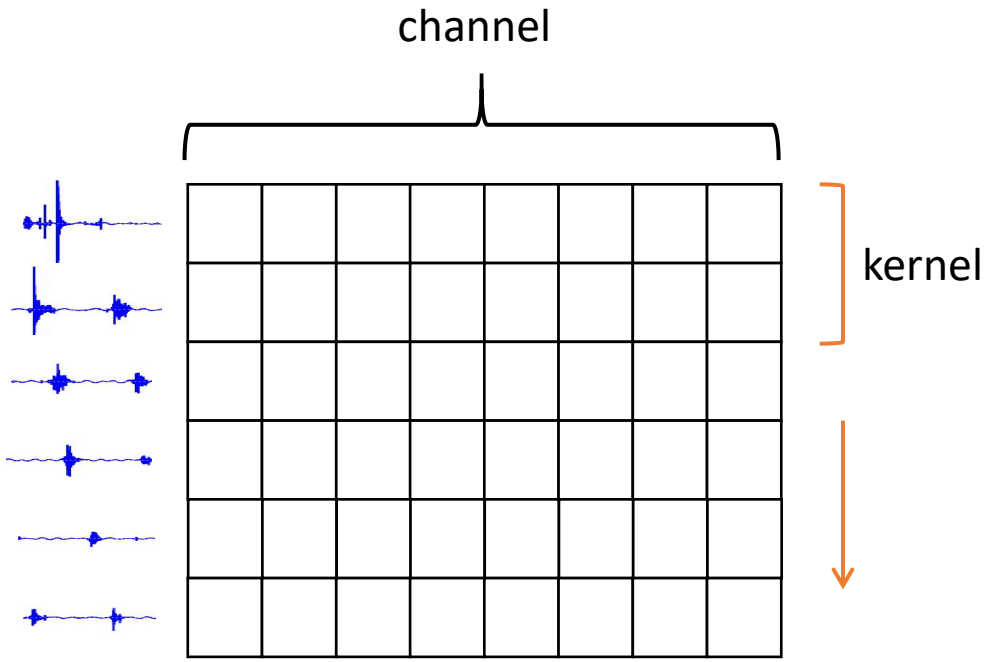
Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean
97.22	97.35	98.47	98.19	97.13	97.67

5.2 demo——语音识别

（1）数据使用完整的AISHELL-1

（2）Encoder(Conv1D)

（3）CTCLoss



5.2 demo——语音识别

（4）字错误率CER为14.6%

（5）语音识别演示

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

语音文字识别和说话人识别

小组成员：王虔翔 王凯 梁晨 和昕 钱鹏屿

2021.9.27