An Overview of Hybrid Automatic Speech Recognition

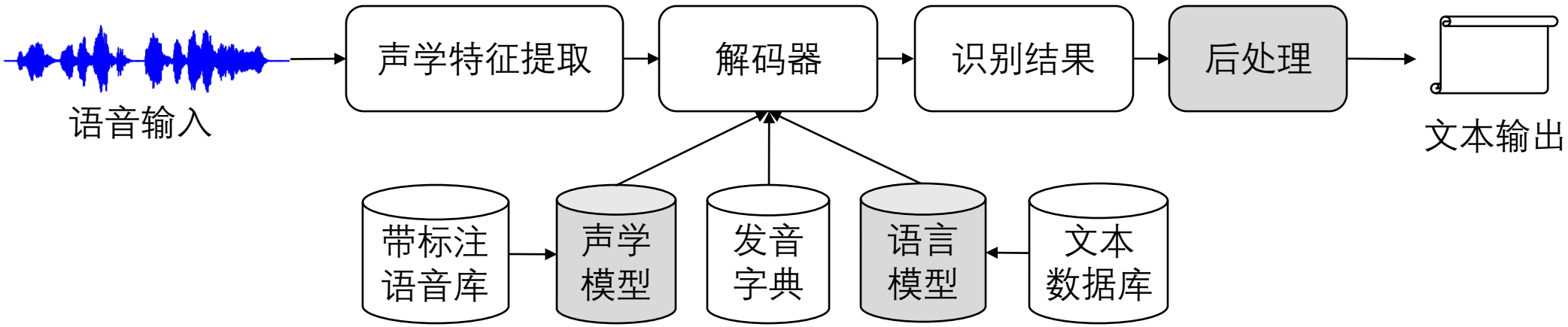
Acoustic Models, Language Models, and Post-Processing

2020-09-17

目录

- 声学模型
  - 基本思路
  - 网络结构
  - 区分性训练
  - LF-MMI
  - 说话人适应
  - 数据增广
  - 弱监督学习
  - 半监督学习
  - 迁移学习
- 语言模型
  - 数据增广与筛选
  - 模型融合方法
  - 重打分/二次解码
- 后处理
  - 系统结果融合
  - 文本检错纠错
  - 标点符号恢复

语音识别系统结构



传统语音识别系统结构示意图

语音识别概率模型

- 语音识别任务目标：**O(observation)**是输入声学特征观测序列，**W**是字/词序列

$$\hat{W} = \arg\max_{W}\{P(W|O)\}$$

- **贝叶斯定理**

声学模型 语言模型

$$\hat{W} = \arg\max_{W}\{\,P(W|O)\,\} = \arg\max_{W}\{\,\frac{P(X|O)\,P(W)}{P(O)}\,\} = \arg\max_{W}\{\,P(O|W)\,P(W)\,\}$$

- **音素**：作为声学建模的基本单元，L表示单词W对应的发音音素序列，由**发音词典**给出

$$P(O|W) = \sum_{L} P(O|L)P(L|W)$$

- **三音子**：考虑音素出现的上下文位置，C是三音素

$$P(O|L) = \sum_{C} P(O|C)\,P(C|L)$$

- **三音子状态共享(senone)：**三音素建模导致概率估计很稀疏，使用决策树聚类

$$P(O|C) = \sum_{S} P(O|S)P(S|C)$$

声学模型/基本思路

$$\hat{W} = \arg\max_W \{P(W|O)\}$$

贝叶斯公式

$$\hat{W} = \arg\max_W \{P(O|W)\,P(W)\}$$

发音词典建模

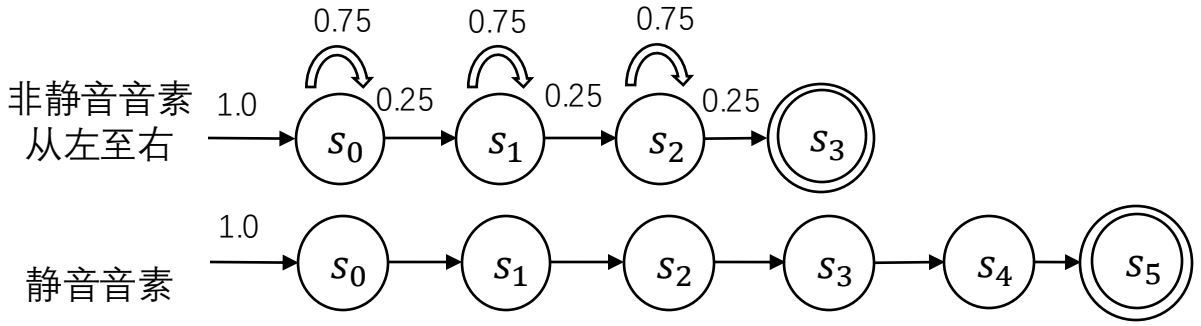
$$P(O|W) = \sum_L P(O|L)P(L|W)$$

三音子模型

$$P(O|L) = \sum_C P(O|C)P(C|L)$$

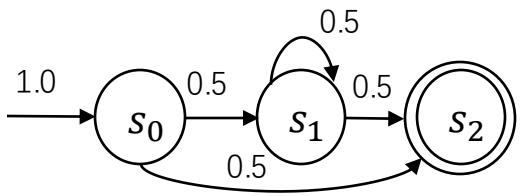
决策树状态聚类

$$P(O|C) = \sum_S \boldsymbol{P(O|S)}P(S|C)$$



各状态之间可以相互转移，并且包含自环

chain model



$\mathbf{P(O|S)}$

(观测)声学特征矢量**序列O**到隐含状态**序列S**的似然度

$$P(O|S) = a_{s_o s_1} \prod_{t=1}^{T} a_{s_t s_{t+1}} b(o_t|s_t)$$

- 各状态之间的转移概率a_{ij}： **HMM**建模
- 发射概率/观测矩阵 $b(o_t|s_i)$ 由GMM / **DNN**来建模

$$b(o_t|s_i) = \frac{P(s_i|o_t)P(o_t)}{P(s_i)}$$

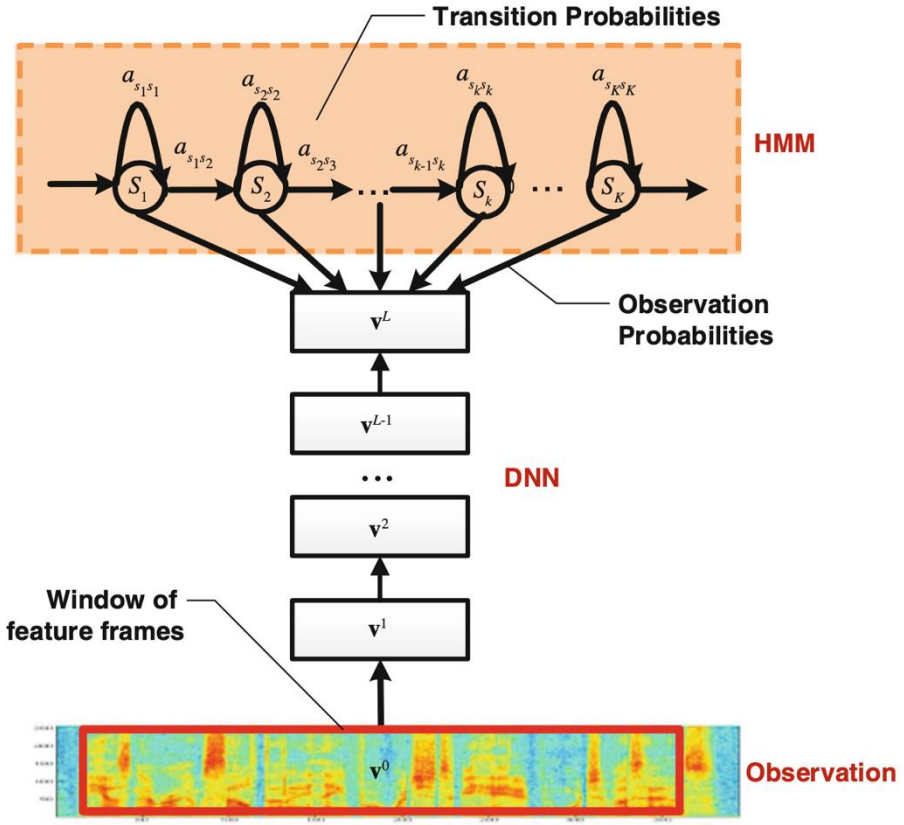
- $P(o_t)$：每帧语音特征的出现概率
- $P(s_i)$：不同状态的先验概率，可以从训练数据的强制对齐结果中统计得到
- $\boldsymbol{P(s_i|o_t)}$：给定声学特征下对应不同状态的概率

↓

转换为分类问题：根据声学特征 预测 聚类状态

声学模型/网络结构

- DNN



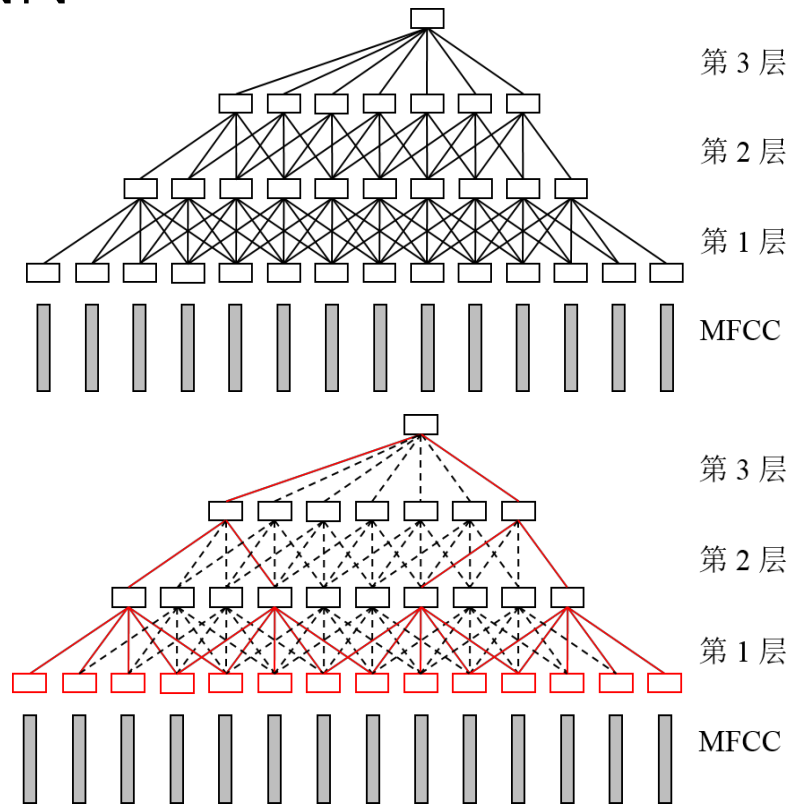
- 各三音子状态之间的转移概率a_{ij}：**HMM**建模
- 发射概率/观测矩阵 $b(o_t|s_i)$ 由GMM/**DNN**来建模

$$b(o_t|s_i) = \frac{P(s_i|o_t)P(o_t)}{P(s_i)}$$

- $\boldsymbol{P(s_i|o_t)}$：给定声学特征下对应不同状态的概率
- DNN相当于一个**分类器**
  - 训练数据：GMM/DNN声学模型**强制对齐**的结果
  - 输出层类别数：聚类后的状态数
  - 帧级别目标函数：交叉熵 cross-entropy
- 将DNN-HMM中的DNN视为分类器之后
  - RNN系列，CNN，TDNN-*，FSMN …

声学模型/网络结构

- TDNN



TDNN的特点

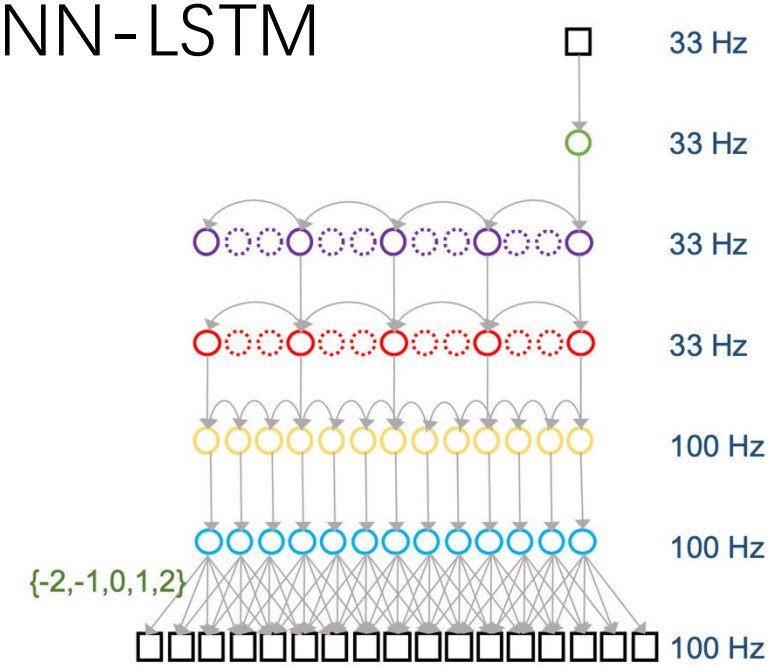
- 与DNN全连接不同，只关注前一层上下文时间范围的信息

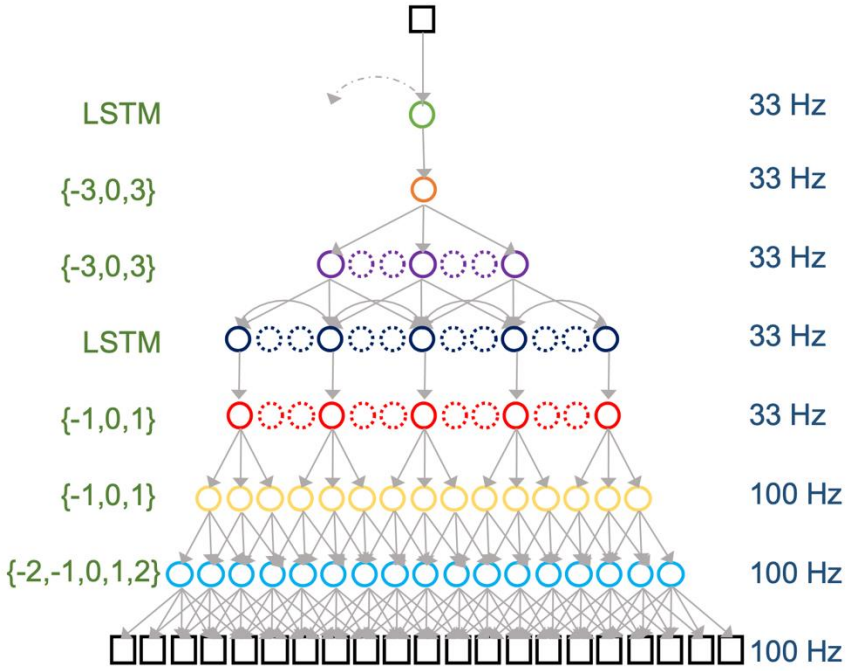
TDNN在声学建模任务中的改进

- 降采样
  - 原因：相邻的隐含层节点输入的上下文context是高度重复和相关的
  - 越深层的节点能够学习覆盖到更长时间范围内的语音特征信息
  - 随着深度的加深，frame rate也在变低
  - 浅层捕捉短期语音特征变化
  - 加速训练5倍左右
- 非对称的context输入
  - 左边context比右边长，降低latency
  - 实验中WER更低

声学模型/网络结构

- TDNN-LSTM

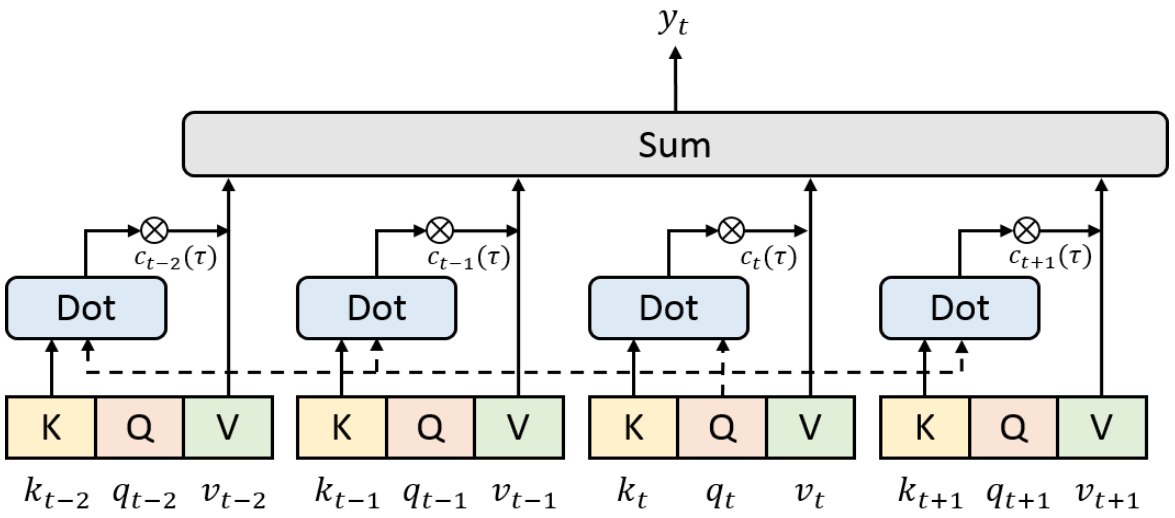


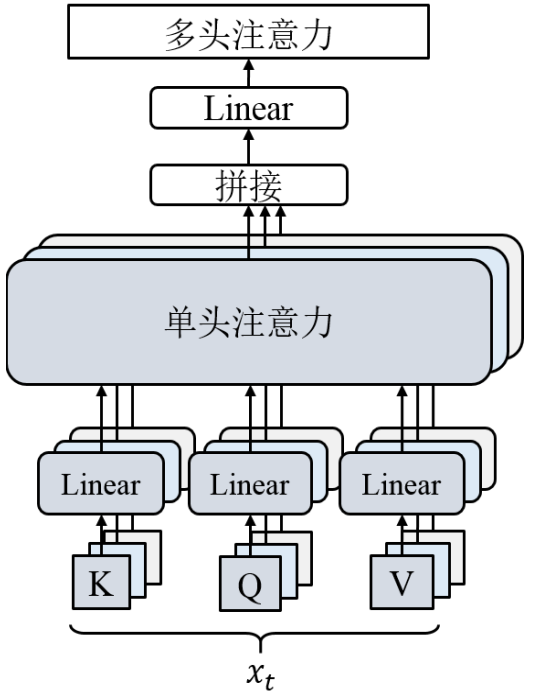


- TDNN-LSTM > Bi-LSTM > LSTM
- TDNN与LSTM交叉堆叠是最优的模型结构
- 输入层没有降采样，frame rate与语音信号相同，之后的网络层会降低frame rate

声学模型/网络结构

- TDNN-Attention





声学模型/网络结构

- TDNN-Attention

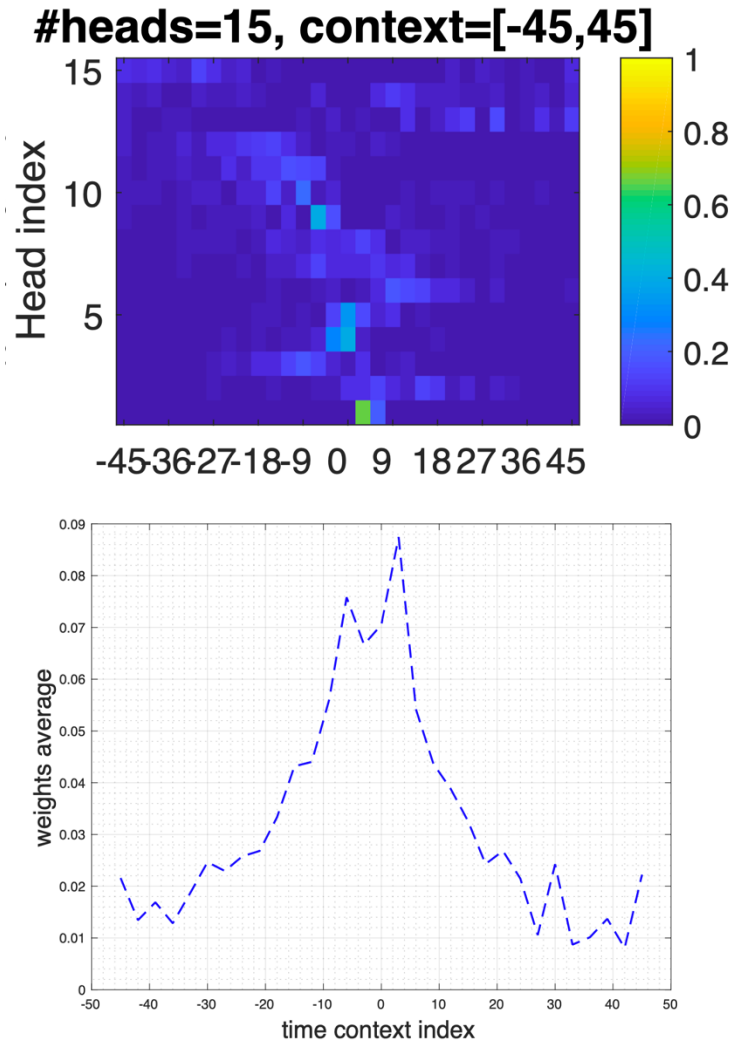
- Time-Restricted 限制上下文范围

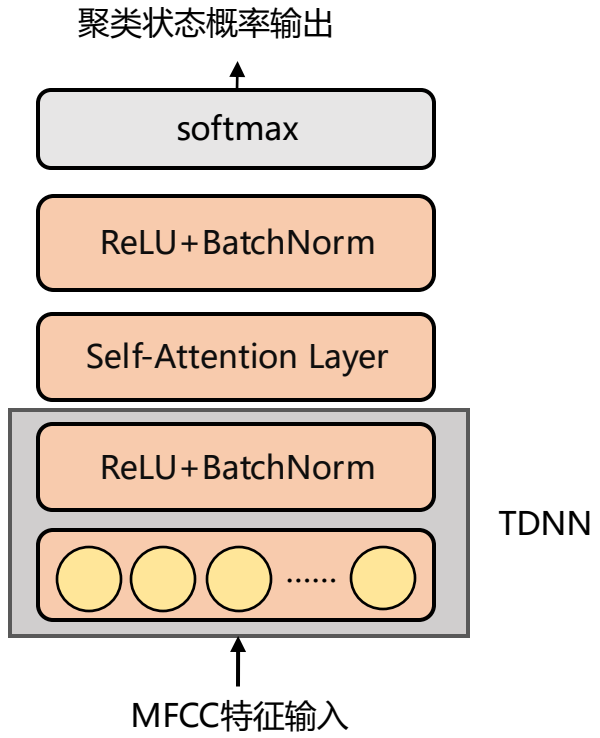
$$y_t = \sum_{\tau=t-L}^{t+R} c_t(\tau) v_t$$

$$c_t(\tau) = \frac{\exp(q_t \cdot k_\tau)}{Z_t}$$

- 位置编码(sin→独热向量)
  - L+1+R维

- Attention层的位置
  - 靠近输出层位置



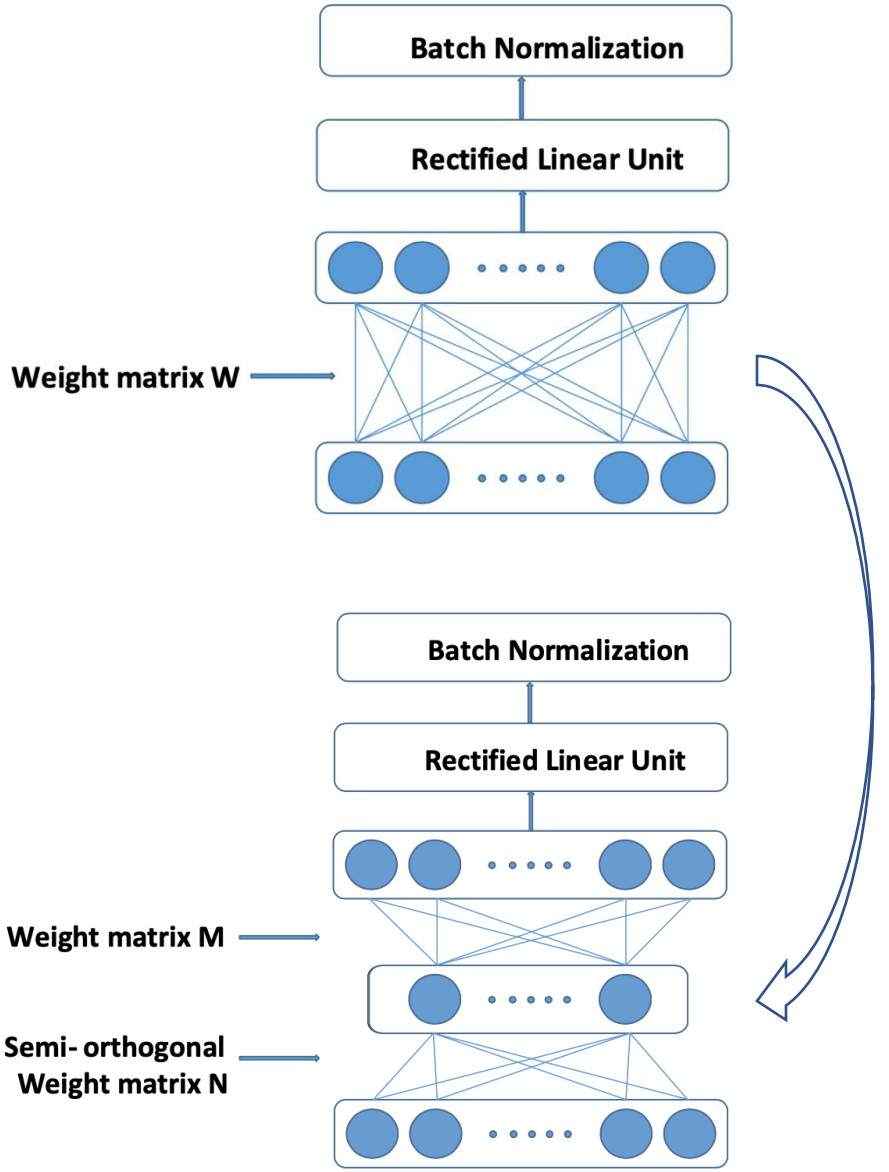


声学模型/网络结构

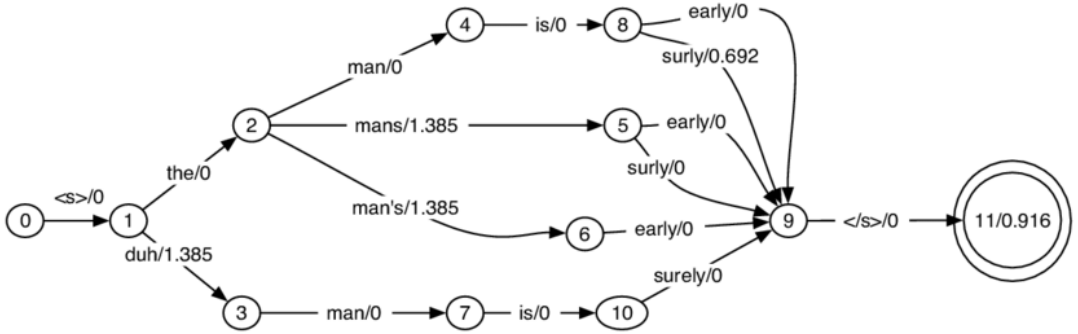
- TDNN-F
  - 权重矩阵分解：$\mathrm{W} = \mathrm{MN}$
  - 半正交约束：$\mathrm{MM}^{\mathrm{T}} = \mathrm{P} \rightarrow \mathrm{I}$
  - 定义：$\mathrm{Q} = \mathrm{P} - \mathrm{I}$
  - 优化目标函数：$f = \mathrm{tr}(\mathrm{QQ}^{\mathrm{T}}) \rightarrow 0$
  - 梯度下降求解：

$$\begin{aligned}\mathrm{M}^{*} &= \mathrm{M} - \lambda \frac{\partial f}{\partial M} \\ &= \mathrm{M} - \lambda \times 4\mathrm{QM} \\ &= \mathrm{M} - 4\lambda(\mathrm{MM}^{\mathrm{T}} - \mathrm{I})\mathrm{M}\end{aligned}$$

  - 其他改进：
    - 跳层连接
    - dropout schedule (对TDNN没用)



声学模型/区分性训练



- MLE → MMI (Maximum Mutual Information)
  - $b(o_t|s_t) = \frac{P(s_t|o_t)P(x_t)}{P(s_t)} \rightarrow \log P(o_t|s_t) = \log P(s_t|o_t) - \log P(s_t)$ pseudo log-likelihood
  - 令$y_{ut} = P(s_t|o_t)$，则DNN部分的优化目标为：

$$\mathcal{F}_{CE} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u}\log y_{ut}(s_{ut})$$

$$\theta_{ML} = \arg\max_{\theta}\{P_\theta(O|W)\}$$

$$\theta_{\mathrm{ML}} = \arg\max_{\theta}\sum_{u}\log P_\theta(O_u|W_u)$$

$$\theta_{MMI} = \arg\max_{\theta}\{P_\theta(W|O)\}$$

$$\theta_{\mathrm{MMI}} = \arg\max_{\theta}\sum_{u}\log P_\theta(W_u|O_u)$$

$$= \arg\max_{\theta}\sum_{u}\log\frac{P_\theta(O_u|W_u)P(W_u)}{P(O_u)}$$

$$= \arg\max_{\theta}\sum_{u}\log\frac{P_\theta(O_u|W_u)P(W_u)}{\sum_W P_\theta(O_u|W)\,P(W)}$$

$$\mathcal{F}_{MMI} = \sum_{u}\log\frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)}$$

κ是acoustic scaling factor

Lattice-based MMI

- 分母：不仅与正确标注有关，还和所有可能的词序列W有关
- 解决方法：用解码生成的Lattice来近似所有可能的词序列W(分母求和)
- 在已有的声学模型的基础上，修改目标函数，利用解码结果基于新的目标函数训练模型参数

声学模型/区分性训练

- MMI → bMMI (boosted MMI)

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)} \qquad \mathcal{F}_{BMMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W) e^{-b\,A(W,W_u)}}$$

$A(W, W_u)$ 表示两个序列间的相似性，或者W与W_u相比的准确率
b是增强系数，相当于给准确率低的W更大的权重

- MPE (Minimum Phone Error) / sMBR (state-level Minimum Bayes Risk)

$$\mathcal{F}_{MBR} = \sum_u \frac{\sum_W p(\mathbf{O}_u|S)^\kappa P(W) A(W, W_u)}{\sum_{W'} p(\mathbf{O}_u|S)^\kappa P(W')}$$

MPE： phone-level
sMBR： state-level

声学模型/LF-MMI

- LF-MMI (chain)

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)}$$

  - 分母部分：从MMI的词级别转换为音素级别
    - phone-level LM：训练集**phone对齐**结果训练N-gram音素级语言模型
    - 不需要对全部训练语料进行解码获取Lattice，所以是Lattice-Free(LF)
    - 合成音素级别的解码图HCLG.fst → HCP (P是音素级语言模型)
  - 分子部分：仍然使用Lattice结果(GMM-HMM/或DNN-HMM得到)

声学模型

- 实验结果对比

		TDNN	TDNN+Attention
TED-LIUM	dev	8.6	**8.3**
	test	8.9	**8.7**
AMI-SDM	eval	43.7	**43.3**
	dev	39.9	**39.5**
AMI-IHM	eval	21.4	**21.1**
	dev	21.4	**21.1**

* fullset/callhm/swbd

		TDNN-LSTM	TDNN-LSTM+ Attention
TED-LIUM	dev	8.3	**8.0**
	test	8.5	**8.2**
AMI-SDM	eval	42	**41.4**
	dev	38.6	**38.2**
AMI-IHM	eval	21.1	**20.5**
	dev	21.1	**20.9**

* fullset/callhm/swbd

Switchboard	Params	Eval2000		RT03	RTF
		SWBD	Total		
TDNN	19M	9.5	14.3	17.5	0.36
BLSTM	41M	9.2	13.7	16.0	1.77
TDNN-LSTM	40M	9.0	13.5	15.6	1.26
TDNN-F	23M	**8.7**	**12.7**	**15.1**	0.45

Table 3: Results (% WER) of the DNNs trained on the full 300 hour training set using different criteria.

	Hub5 '00			Hub5 '01			
System	SWB	CHE	Total	SWB	SWB2P3	SWB-Cell	Total
GMM BMMI	18.6	33.0	25.8	18.9	24.5	30.1	24.6
DNN CE	14.2	25.7	20.0	14.5	19.0	25.3	19.8
DNN MMI	12.9	24.6	18.8	13.3	17.8	23.7	18.4
DNN sMBR	12.6	24.1	18.4	13.0	17.7	22.9	18.0
DNN MPE	12.9	24.1	18.5	13.2	17.7	23.4	18.2
DNN BMMI	12.9	24.5	18.7	13.2	17.8	23.5	18.3

Table 4: Performance of LF-MMI on various LVCSR tasks with different amount of training data, using TDNN acoustic models

Database	Size	WER		
		CE	CE → sMBR	LF-MMI
AMI-IHM	80 hrs	25.1	23.8	22.4†
AMI-SDM	80 hrs	50.9	48.9	46.1†
TED-LIUM	118 hrs	12.1	11.3	11.2*
Switchboard	300 hrs	18.2	16.9	15.5
Librispeech	1000 hrs	4.97	4.56	4.28
Fisher + SWBD	2100 hrs	15.4	14.5	13.3

声学模型/说话人适应

- X-vector：说话人特征向量
  - 目标：将一段不定长语音映射成一个特征向量
  - 沿时间方向将各帧声学特征的统计量进行拼接
    - 统计量：均值、标准差

Layer	Layer context	Total context	Input x output
frame1	$[t-2, t+2]$	5	120x512
frame2	$\{t-2, t, t+2\}$	9	1536x512
frame3	$\{t-3, t, t+3\}$	15	1536x512
frame4	$\{t\}$	15	512x512
frame5	$\{t\}$	15	512x1500
stats pooling	$[0, T)$	T	1500Tx3000
segment6	$\{0\}$	T	3000x512
segment7	$\{0\}$	T	512x512
softmax	$\{0\}$	T	512xN

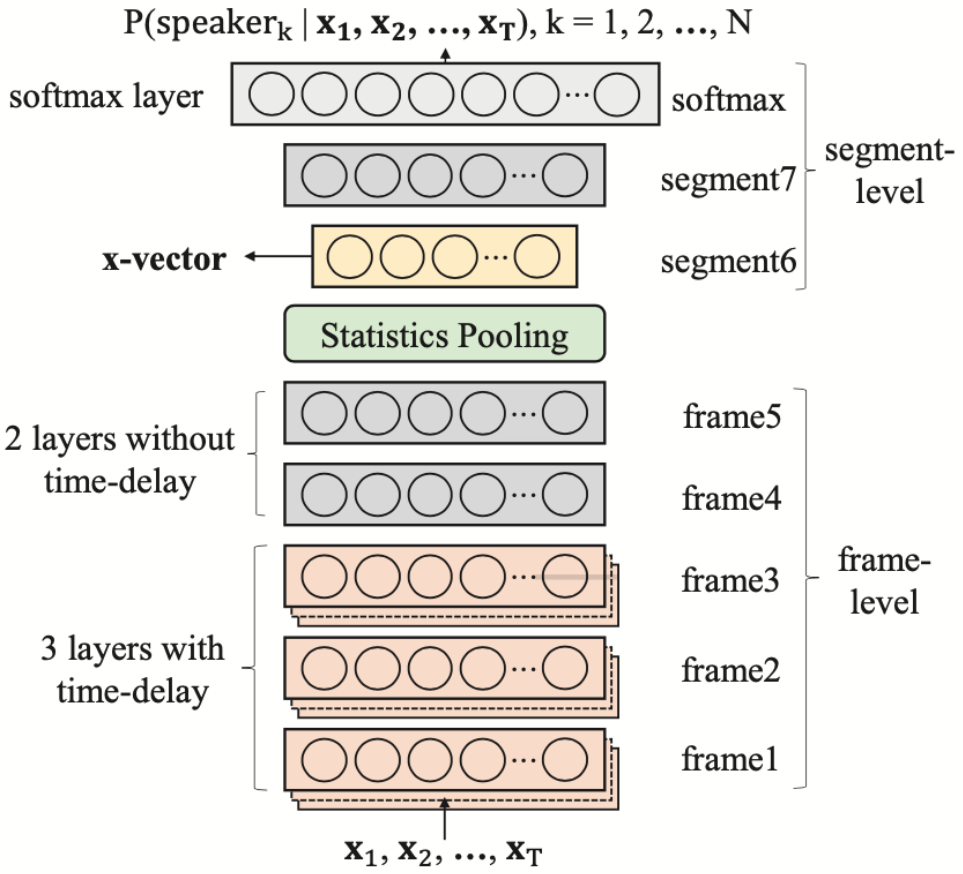


图2 x-vector网络结构

声学模型/说话人适应

- 目标：使得声学模型能够兼顾到说话人的特性，提高识别效果
- 说话人适应 v.s. 说话人识别
  - fMLLR：说话人适应 → 说话人识别
  - i-vector：说话人识别 → 说话人适应
  - x-vector：说话人识别 → 说话人适应?
- 基于X-vector的说话人适应
  - 借鉴i-vector在说话人适应中的方法
    - 训练数据筛选：去除时长短和语音片段数少的说话人
    - 细分说话人标签，增加说话人丰富性
      - 将语音片段数多的说话人拆分成多个说话人
    - 与mfcc/fbank特征拼接作为声学模型的输入
  - 去除说话人识别中的LDA降维模块
    - 类似于Bottleneck，直接将神经网络中提取所需维度的特征

声学模型/说话人适应

- X-vector：说话人特征向量
  - 模型的训练目标：说话人多分类
  - 说话人识别的后端：LDA降维 + PLDA分类器
  - 提高鲁棒性：数据增强
    - MUSAN噪声库：babble music noise
    - RIRS_NOISES混响库：reverb
  - 相比于i-vector的优势
    - 在说话人识别任务中性能更优
    - 数据增强后提升效果相比于i-vector更明显
    - 适合大数据量级的说话人识别模型训练

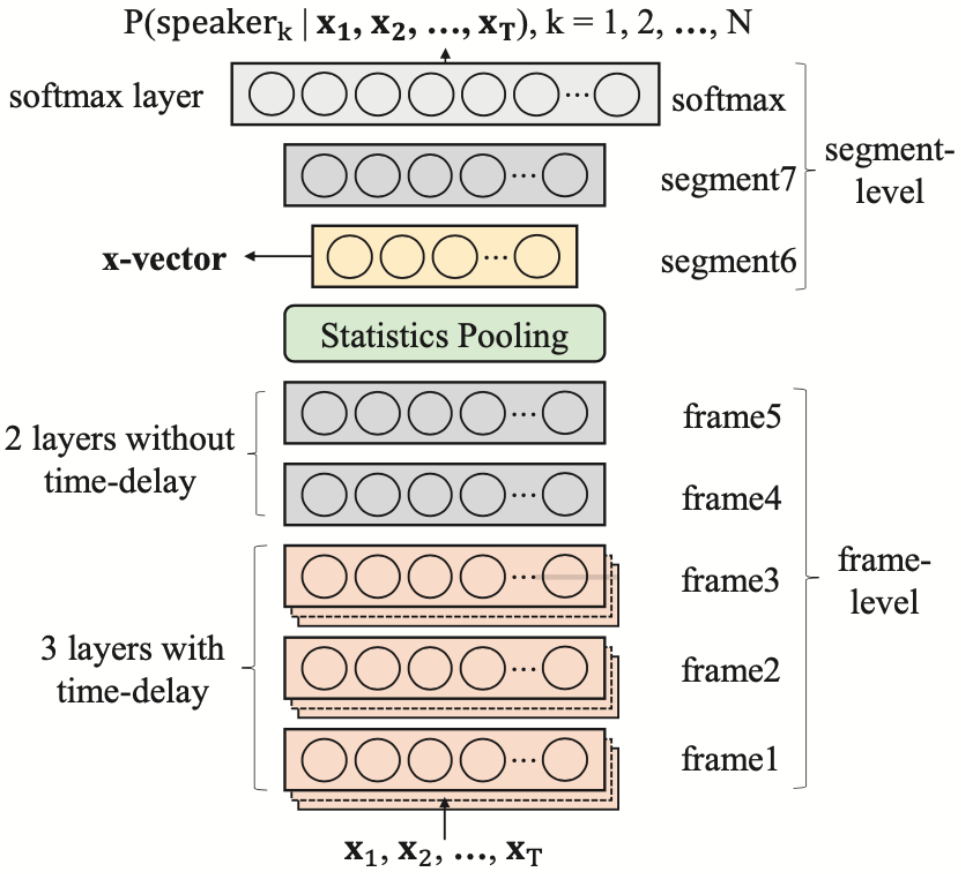


图2 x-vector网络结构

声学模型/说话人适应

- 基于X-vector的说话人适应性训练
  - 最终实验方案（TED_LIUM r3数据）
    - 说话人拆分
    - x-vector选用200维
    - 200维X-vector直接从DNN中提取，不用LDA
  - i-vector与x-vector在结果层面的融合
    - ROVER融合算法
  - 其他优化思路（未实验验证）
    - 数据增强：加噪、加混响
    - Statistics Pooling增加均值、方差之外的统计量

d	Speaker Modification	WER		WER 4-gram		WER RNNLM	
		Dev	Test	Dev	Test	Dev	Test
100	No	8.37	8.53	7.82	8.13	6.69	7.07
	Yes	8.44	8.28	7.83	7.94	6.76	6.97
200	No	8.18	8.40	7.73	7.95	6.49	6.94
	Yes	8.29	8.36	7.74	7.89	6.48	**6.78**

表1 说话人切分后的实验结果对比

d	Extraction Strategy	WER		WER 4-gram		WER RNNLM	
		Dev	Test	Dev	Test	Dev	Test
100	LDA	8.52	8.48	7.87	8.05	6.69	7.15
	no LDA	8.44	8.28	7.83	7.94	6.76	6.97
200	LDA	8.31	8.57	7.78	8.01	6.65	6.91
	no LDA	8.29	8.36	7.74	7.89	6.48	6.78

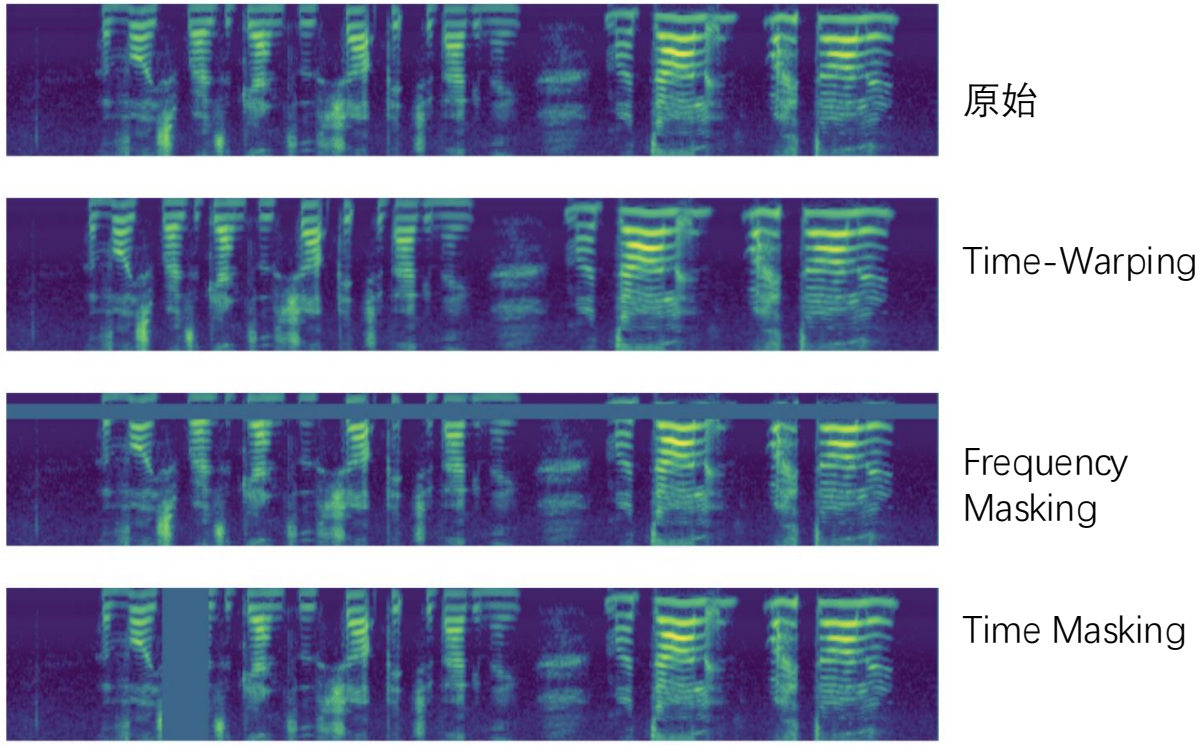
表2 是否有LDA对实验结果的影响

system	WER		WER 4-gram		WER RNNLM	
	Dev	Test	Dev	Test	Dev	Test
i-vector	7.85	8.39	7.22	7.76	6.20	6.95
x-vector	8.29	8.36	7.74	7.89	6.48	**6.78**
feature fusion	8.10	8.36	7.46	7.80	6.40	6.90
system fusion	7.70	8.28	7.19	7.71	**6.06**	**6.71**

表3 i-vector与x-vector的融合

声学模型/数据增广

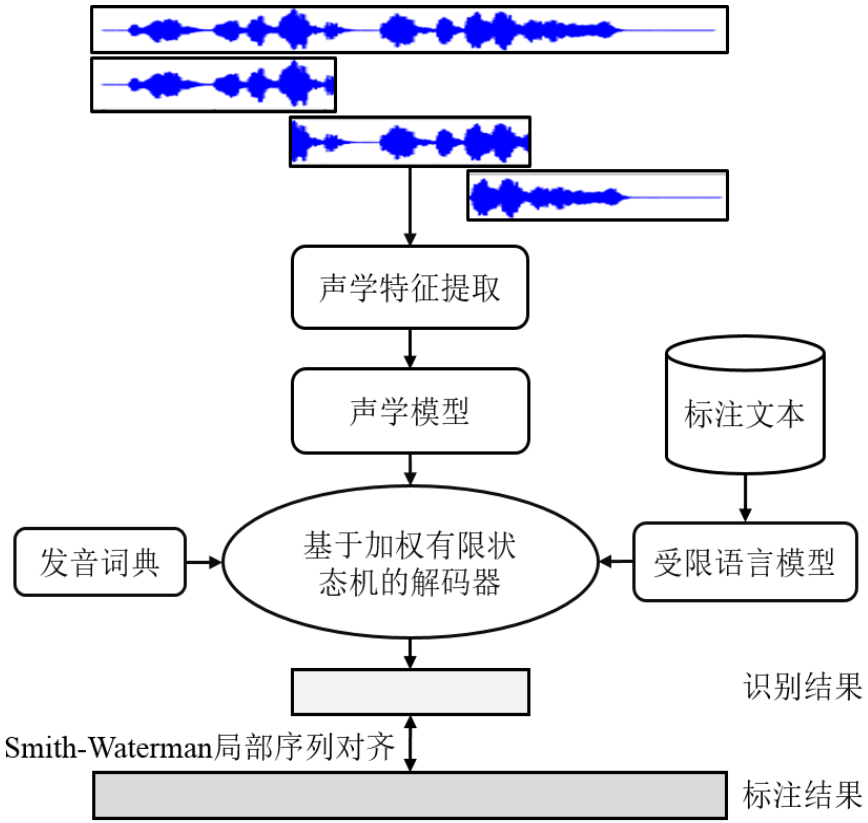
- 数据增广方法总结
  - 速度扰动：0.9、1.0、1.1
  - 音量扰动：[0.125, 2] 随机选取
  - 加性噪声：
    - MUSAN
      - music 音乐声
      - babble 人声
      - noise 噪声
    - RIRS_NOISES：加混响 reverberation
    - 注意信噪比SNR
  - 频谱增强
  - frame-shift (chain)
  - mixup



声学模型/弱监督学习

- 数据清洗
  - 长音频切分和低质量语音处理
    - biased LM/restricted LM 受限语言模型
    - smith-waterman

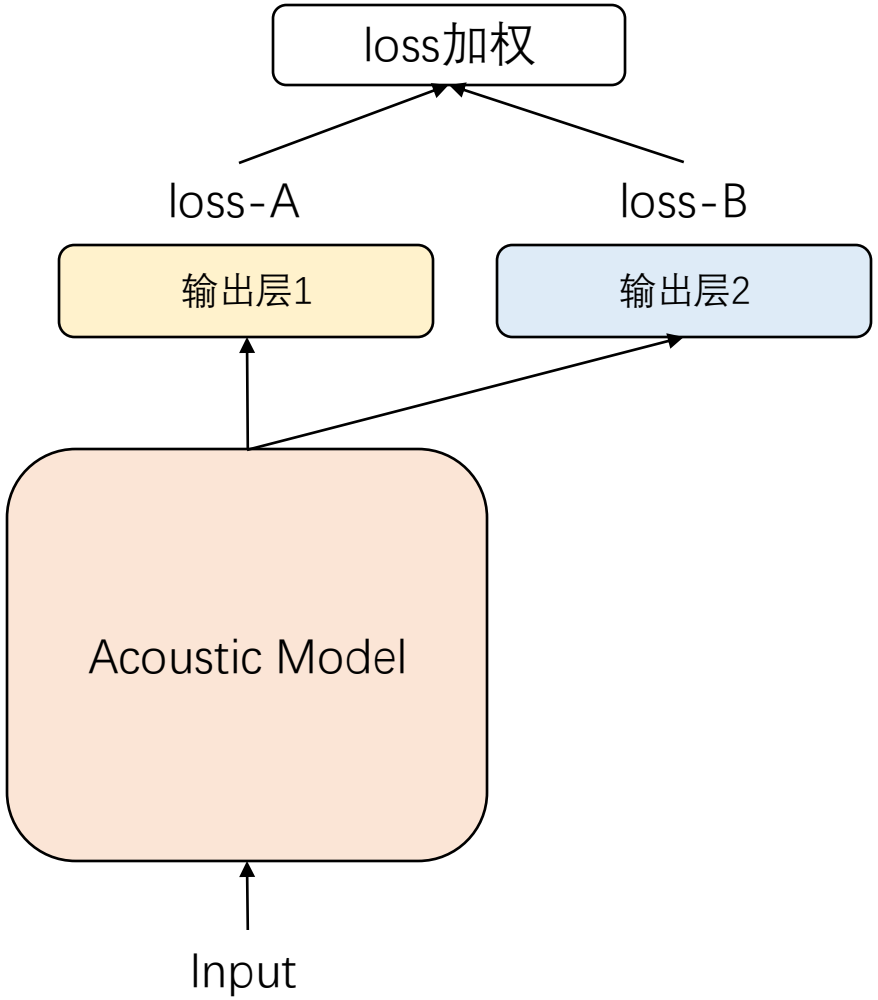
    - 基于滑动窗解码
      - 固定时间窗长或者先做VAD
      - 在时间窗内解码
      - 将解码结果与标注进行局部序列对齐



声学模型/半监督学习

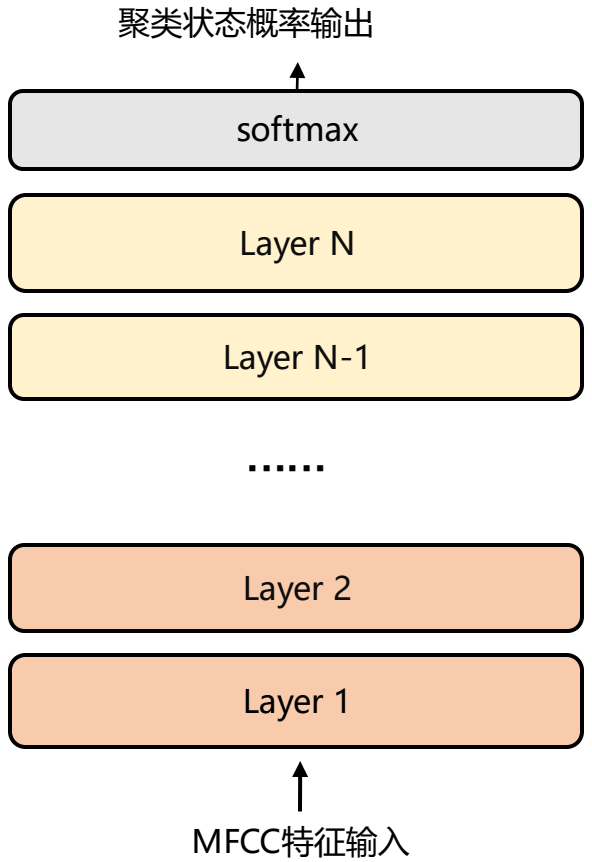
- 伪标注生成：解码结果筛选
  - confidence score
  - lattice entropy

- 多任务multi-task learning
  - 有标注的数据 loss-A
  - 生成了伪标注的数据 loss-B
  - 优化的loss：loss-A和loss-B的加权和
  - 最终使用的有标注的网络部分



声学模型/增量学习+迁移学习

- 策略一：对齐结果融合，继续或重新迭代训练模型
  - 特色：模型结构不变，训练超参数不变
- 策略二：使用小数据集，在已有模型上继续迭代训练fine-tune
  - 特色：模型结构不变，训练超参数改变
  - 方案1：降低所有层的学习率和迭代轮数
  - 方案2：其它网络层参数固定，只更新输出层或最后若干层
- 策略三：靠近输出层增加随机初始化的网络层
  - 特色：模型结构改变，训练超参数改变
  - 新加层和输出层较大学习率，其他层较小学习率



语言模型/数据增广与筛选

- 适用场景：低资源语种或应用场景

- 文本筛选方法
  - 标注文本的数据不足（低资源场景），用集外文本提高语言模型效果
  - 筛选任务的目标：从集外文本中筛选出与训练标注文本相近的文本数据
  - 筛选数据的用途：训练数据扩充、发音词典扩充（结合g2p）

  - 文本相似度 / 相关性分析
    - 困惑度排序
    - 相对交叉熵排序
    - TF-IDF向量相似度
    - Doc2vec向量相似度

语言模型/数据增广与筛选

- 文本筛选方法：困惑度排序
  - 使用集内数据训练的语言模型，对每个集外文本句子求出困惑度（Perplexity，PPL）

$$\mathrm{PPL(W)} = \mathrm{P}(w_1 w_2 w_3 \cdots w_N)^{-\frac{1}{N}}$$

$$\mathrm{PPL(W)} = \left[\prod_{i=1}^{N} P(w_i | w_1 w_2 \cdots w_{i-1})\right]^{-\frac{1}{N}}$$

$$\text{2gram：}\ \mathrm{PPL(W)} = \left[\prod_{i=1}^{N} P(w_i | w_{i-1})\right]^{-\frac{1}{N}}$$

  - 按照困惑度对集外文本排序，选出困惑度低于某一阈值的句子

语言模型/数据增广与筛选

- 文本筛选方法：相对交叉熵
  - 交叉熵：
$$H(P_{LM}) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{LM}(w_i|w_1,\cdots,w_{i-1})$$
  - 集内文本训练语言模型A，集外文本训练语言模型B
  - 对于集外每一句文本
    - 分别使用语言模型A、B计算交叉熵，作差得到相对交叉熵
  - 按照相对交叉熵排序

语言模型/数据增广与筛选

- 文本筛选方法：TF-IDF

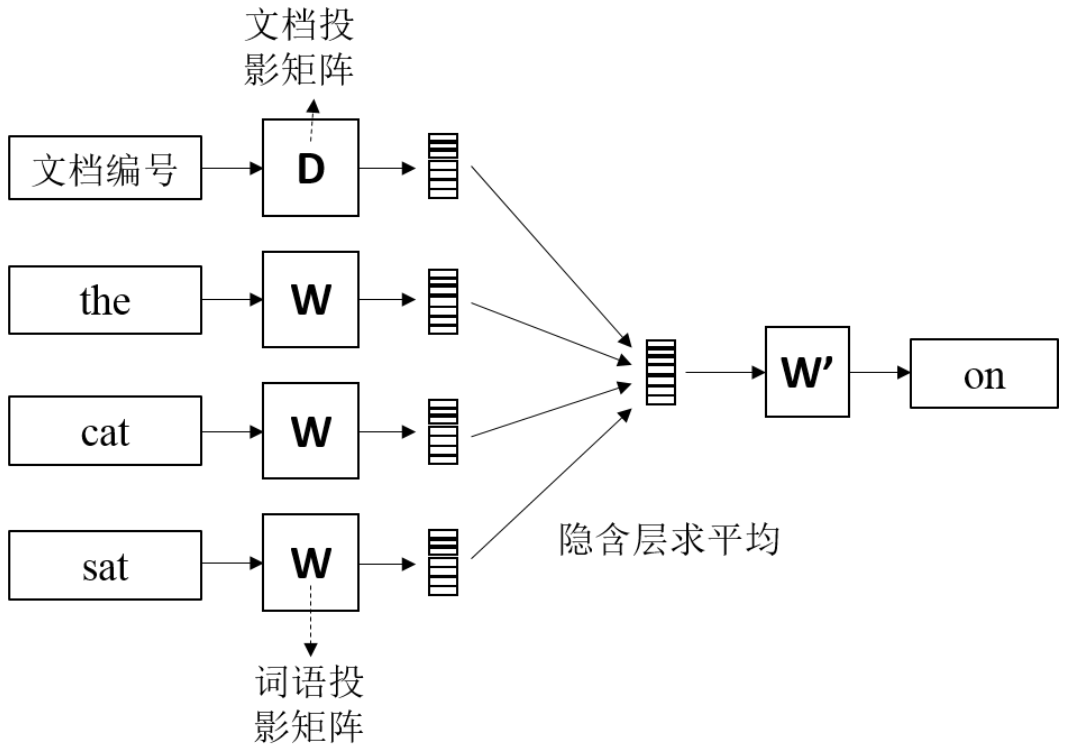
$$\mathrm{TF} = \frac{C(W|D)}{C(D)}$$

$$\mathrm{IDF} = \log \frac{C(M)}{C(M|W)+1}$$

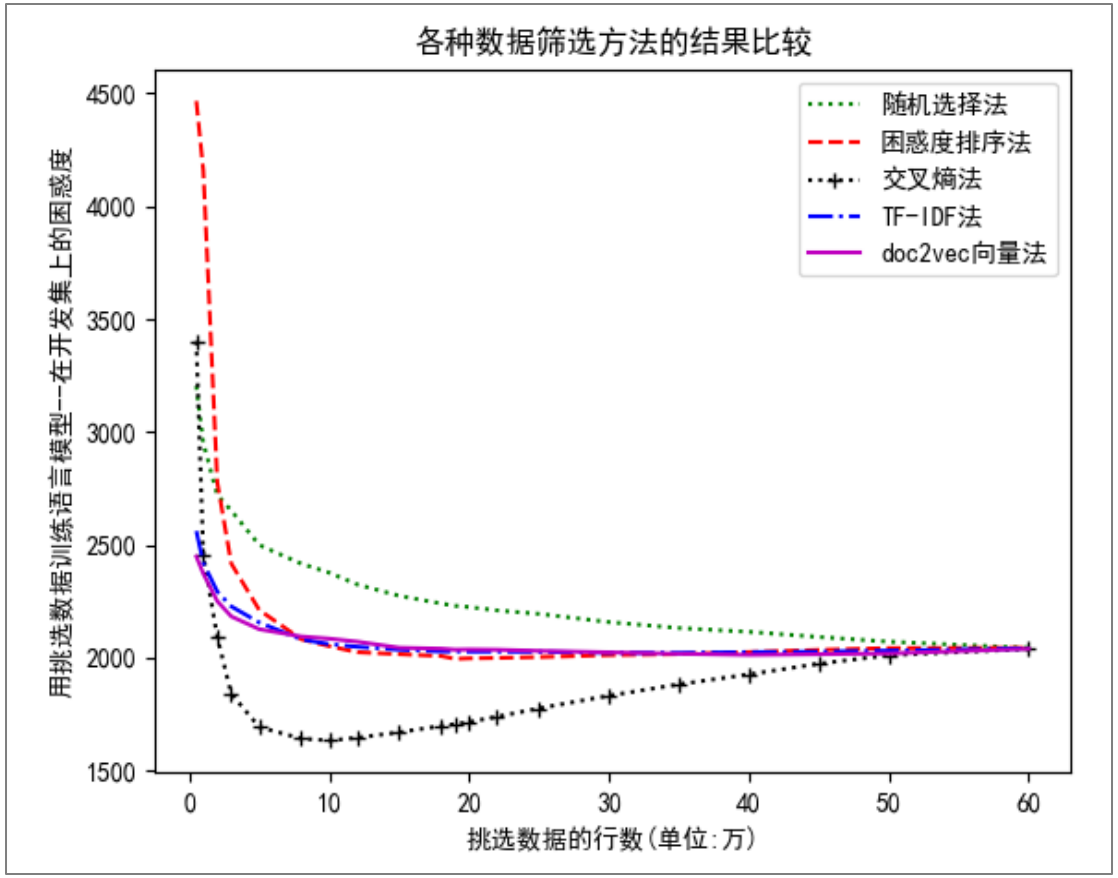
  - 某词的TF-IDF值越大，该词对文章重要性越高，对于文档越关键(关键词)
  - 每篇文章各取出若干个关键词，合并成一个集合
  - 对于每个文档，计算关键词集合中每个词的TF-IDF值，合为一个向量
    - 计算文档之间向量的相似度，值越大表示越相似。

语言模型/数据增广与筛选

- 文本筛选方法：doc2vec
  - Distributed Memory Model
  - 将文档投影为向量，参与到训练中
    - 文档向量在每个CBOW预测中都起到了作用
  - 文本筛选流程
    - 集内文本对应一个文档向量
    - 将集外文档向量与集内文档向量求相似度
    - 根据相似度进行排序



语言模型/数据增广与筛选



语言模型/模型融合方法

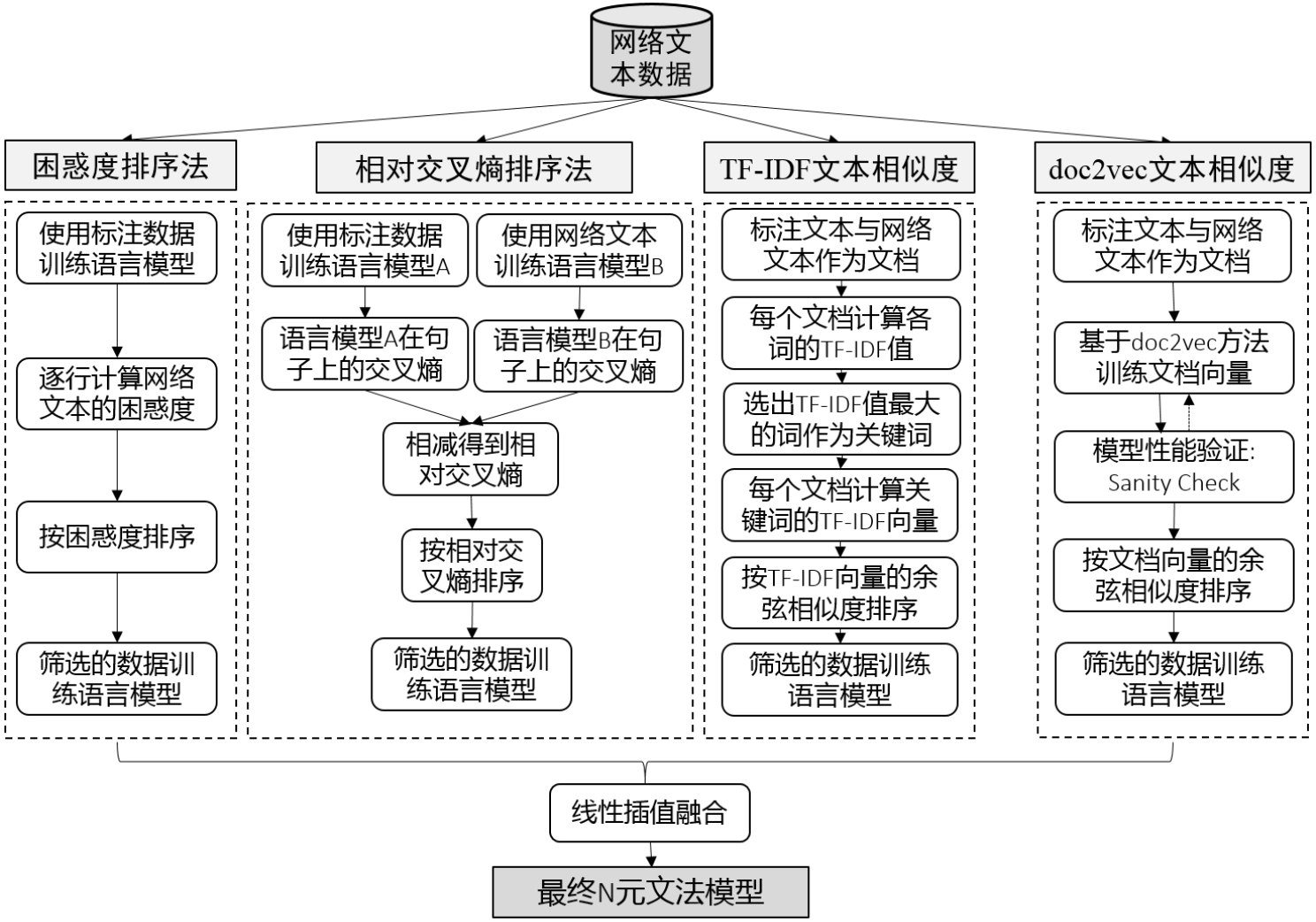
- 如何使用筛选后的数据?

方案一：与集内文本数据融合，训练语言模型
方案二：分别训练语言模型，采用线性插值方法，插值系数使得在开发集上PPL最低

数据来源	集内数据	困惑度法	交叉熵法	TF-IDF法	Doc2vec法
数据行数	3.77万	20万	10万	30万	40万
语言模型	ME3	KN4	KN4	KN4	KN4
困惑度	429.30 WER = 47.7%	1994.43	1632.75	2021.46	2012.21
插值系数	0.825	0.054	0.017	0.042	0.062
各模型插值	PPL = 387.32, WER = 46.8 %				

建模方法	最优语言模型	语言模型困惑度
原始训练数据	ME3	429.30
相对交叉熵法筛选的数据	KN4	1632.75
两部分数据合并	KN4	598.42
两部分数据分别训练语言模型后插值	ME3	409.11

语言模型/模型融合方法


网络文本数据
困惑度排序法
使用标注数据训练语言模型
逐行计算网络文本的困惑度
按困惑度排序
筛选的数据训练语言模型
相对交叉熵排序法
使用标注数据训练语言模型A
使用网络文本训练语言模型B
语言模型A在句子上的交叉熵
语言模型B在句子上的交叉熵
相减得到相对交叉熵
按相对交叉熵排序
筛选的数据训练语言模型
TF-IDF文本相似度
标注文本与网络文本作为文档
每个文档计算各词的TF-IDF值
选出TF-IDF值最大的词作为关键词
每个文档计算关键词的TF-IDF向量
按TF-IDF向量的余弦相似度排序
筛选的数据训练语言模型
doc2vec文本相似度
标注文本与网络文本作为文档
基于doc2vec方法训练文档向量
模型性能验证: Sanity Check
按文档向量的余弦相似度排序
筛选的数据训练语言模型
线性插值融合
最终N元文法模型

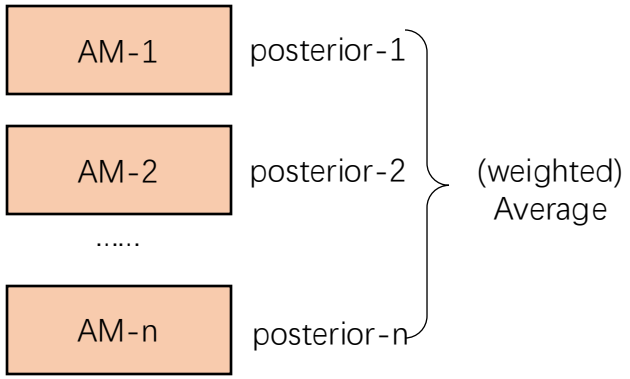
语言模型/重打分

- N-Best/Lattice Rescoring
  - acoustic score + language model score + graph cost
  - 声学模型和graph cost不更改，只修改语言模型分数

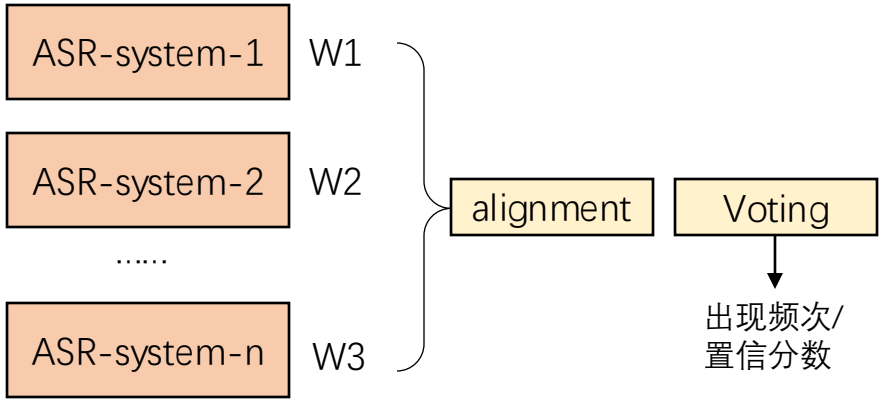
- 重打分模型
  - large n-gram model
  - RNNLM
  - Transformer-LM

后处理/系统融合

- 声学模型：后验概率平均

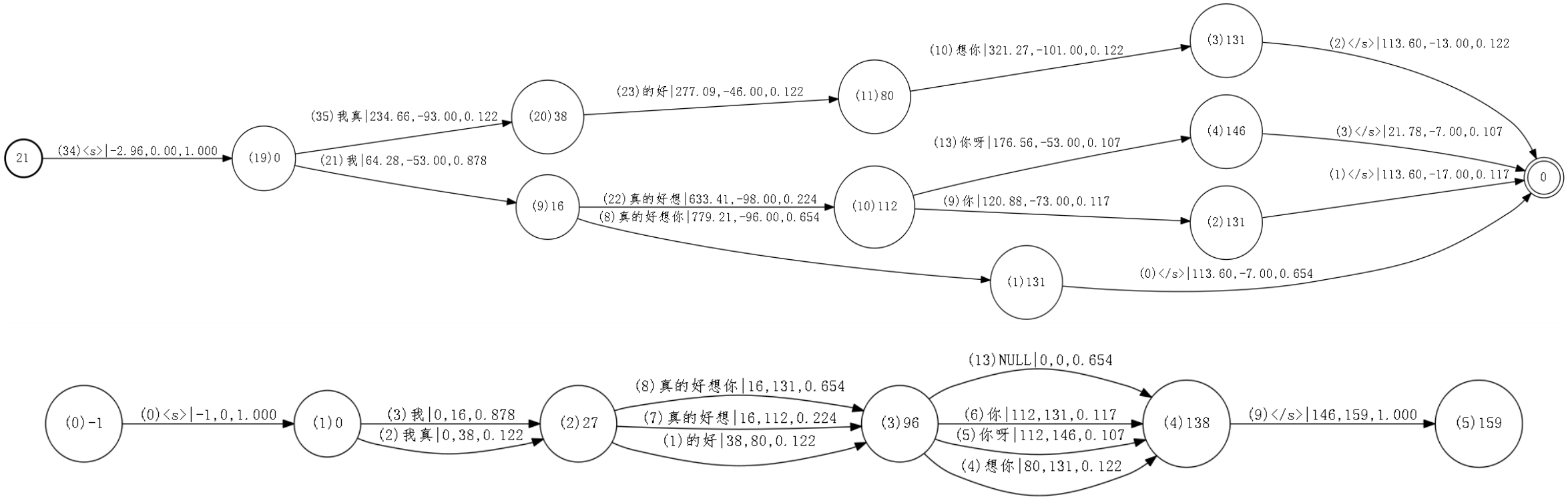


- 输出文本结果：ROVER算法



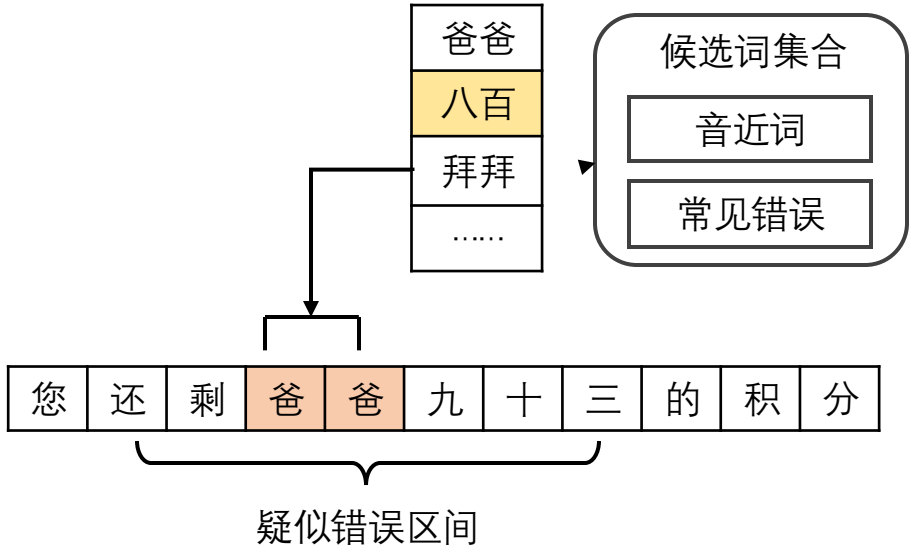
后处理/系统融合

- Lattice：Lattice Combination + MBR Decoding



后处理/文本纠错

- 候选词生成
  - 根据发音音素序列，筛选出音近字词集合
  - 易错词统计，整理成易错词对集合
- 文本纠错
  - 若包含易错词，替换后打分判断是否替代
  - 选择分数最高的候选词作为结果
  - 通过设定阈值调整纠正力度
- N-gram：多组语言模型提高稳健性
  - 字级别 / 词级别
  - 2-gram 3-gram 4-gram

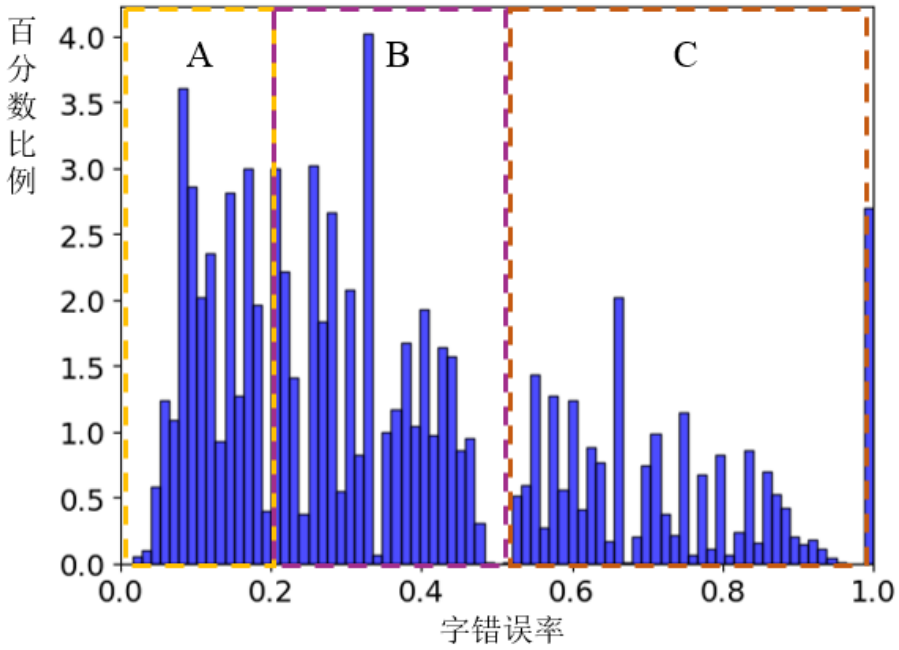


纠错前：	我是保险公司的**带你**小钱友印象吗
纠错后：	我是保险公司的**代理**小钱有印象吗
纠错前：	目前给您订的机票是**男孩**公司明天的航班
纠错后：	目前给您订的机票是**南航**公司明天的航班
纠错前：	**把**电话给您是因为您可以参加积分兑一**反**一的活动
纠错后：	**打**电话给您是因为您可以参加积分兑一**返**一的活动

后处理/文本纠错

- 实验结果

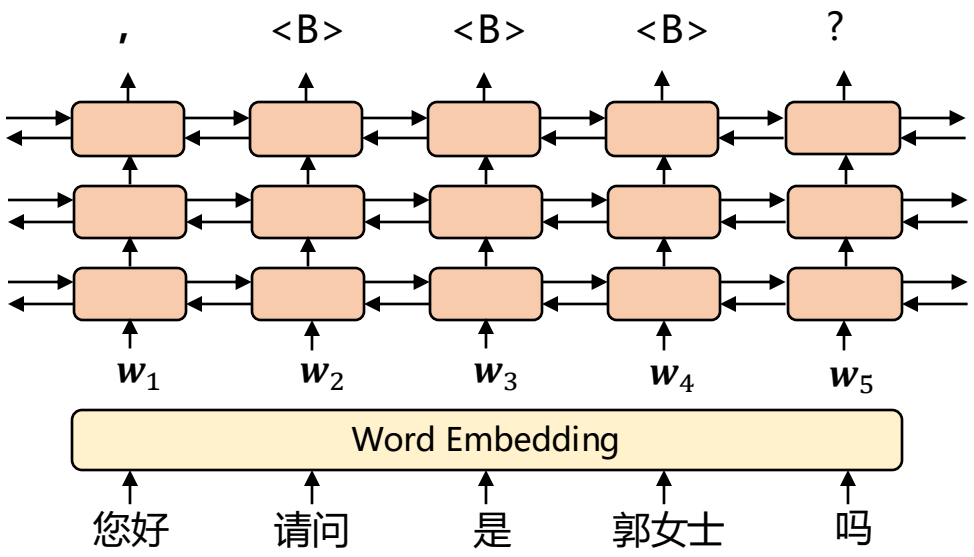
	A组 (0–20%) (错误率较低)		B组 (20–50%) (错误率中等)		C组(50–100%) (错误率较高)	
	纠错前	纠错后	纠错前	纠错后	纠错前	纠错后
替换	6.1%	5.8%	18.1%	19.3%	43.2%	45.8%
插入	3.0%	3.0%	7.1%	7.2%	4.2%	4.4%
删除	2.2%	2.2%	5.4%	5.3%	12.2%	12.2%
合计	11.3%	11.0%	30.6%	31.8%	59.6%	62.4%
相对变化	↓ 2.7%		↑ 3.9%		↑ 4.7%	



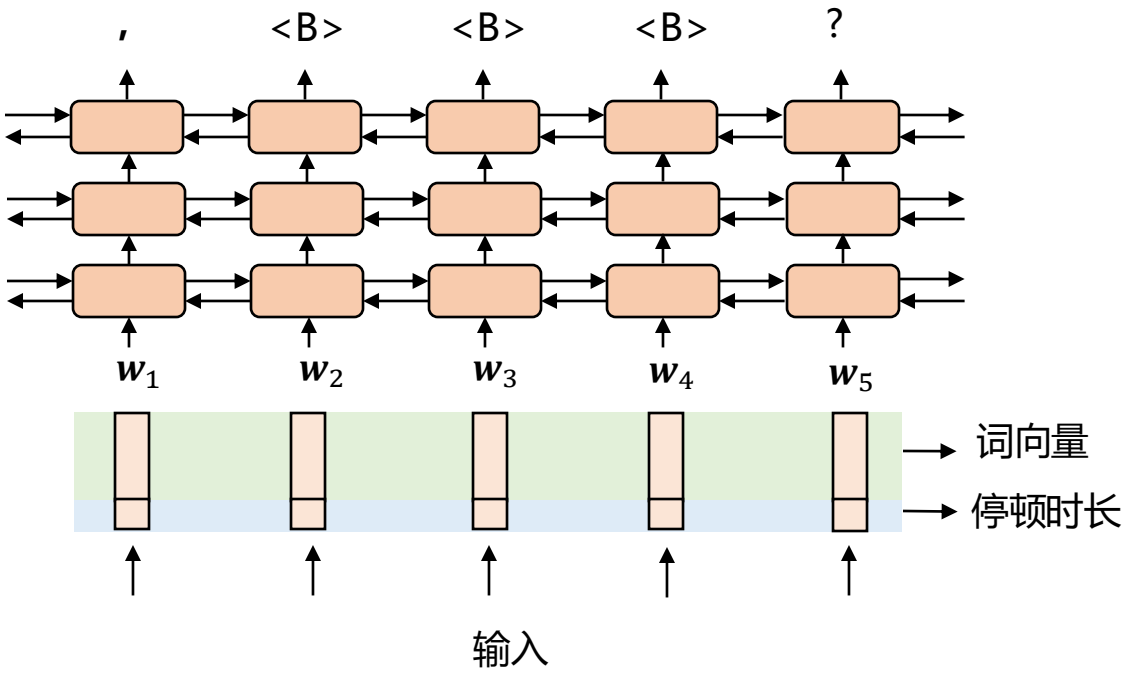
- 结论一：适用于纠正替换类型的错误
- 结论二：适用于错误率较低的识别结果

后处理/标点恢复

- 双向LSTM结构



输入特征加入词之间的停顿时间信息



思考

- 传统Hybrid语音识别的优势
  - 具有SOTA的识别准确率，尤其是数据量较小的场景
  - 模块化的结构更方便进行性能分析与错误诊断
  - 可以只基于语言模型对垂直领域进行场景适应
  - 性能更稳定，经验参数无须过分调参，模型对参数敏感性较低
  - 专业人员维护的Kaldi工具，兼具 前沿算法 & 工程优化
  - Inference时对计算资源的消耗更低，保证CPU下的实时率
  - Kaldi的其他特性
    - 功能相对完备的框架：多通道语音识别、多语种语音识别、说话人识别/分离、语种识别、语音唤醒/关键词检测、VAD 等
    - Lookahead动态解码器：同时节省硬盘占用和内存消耗，适合客户端场景

FST	FST size	WER	RTF
static HCLG	476M HCLG	4.86	0.18
dynamic HCL/G	48M HCL + 77M G	4.86	0.29

    - 更灵活精细的建模控制：发音词典概率建模、静音模型、额外语法grammar修正

思考

- 可能的工作
  - 区分性训练声学模型在口语打分/判错中的效果?
  - 针对弱标注的语音数据（网络音视频）使用数据清洗方法进行数据预处理
  - 针对无标注的语音数据，采用半监督学习方法
  - 从成人声学模型到儿童声学模型，采用迁移学习方法
  - x-vector优化及基于x-vector的online说话人适应（速度和准确率）

  - 语言模型的不同领域适应（模型插值融合）
  - 基于RNNLM/TransformerLM的语言模型重打分（效果提升与时间消耗的关系）

  - 两个或多个模型/系统的结果融合（时间考虑？）

  - 基于Transformer的端到端纠错后处理

参考文献

- Peddinti, Vijayaditya, Daniel Povey, and Sanjeev Khudanpur. "A time delay neural network architecture for efficient modeling of long temporal contexts." *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
- Waibel, Alex, et al. "Phoneme recognition using time-delay neural networks." *IEEE transactions on acoustics, speech, and signal processing* 37.3 (1989): 328-339.
- Peddinti, Vijayaditya, et al. "Low latency acoustic modeling using temporal convolution and LSTMs." *IEEE Signal Processing Letters* 25.3 (2017): 37.
- Povey, Daniel, et al. "A time-restricted self-attention layer for asr." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- Povey, Daniel, et al. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks." *Interspeech*. 2018.
- Leggetter, Christopher J., and Philip C. Woodland. "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models." *Computer speech & language* 9.2 (1995): 171-185.
- Ferras, Marc, et al. "Constrained MLLR for speaker recognition." *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 4. IEEE, 2007.
- Karafiát, Martin, et al. “iVector-based discriminative adaptation for automatic speech recognition." *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011.
- Saon, George, et al. "Speaker adaptation of neural network acoustic models using i-vectors." *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013.
- Peddinti, Vijayaditya, et al. "Reverberation robust acoustic modeling using i-vectors with time delay neural networks." *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

参考文献

- Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- Raj, Desh, et al. "Probing the information encoded in x-vectors." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.
- Bahl, Lalit, et al. "Maximum mutual information estimation of hidden Markov model parameters for speech recognition." *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 11. IEEE, 1986.
- Povey, Daniel. *Discriminative training for large vocabulary speech recognition*. Diss. University of Cambridge, 2005.
- Gibson, Matthew, and Thomas Hain. "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition." *Ninth international conference on spoken language processing*. 2006.
- Povey, Daniel, et al. "Boosted MMI for model and feature-space discriminative training." *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008.
- Veselý, Karel, Arnab Ghoshal, Lukás Burget, and Daniel Povey. "Sequence-discriminative training of deep neural networks." In *Interspeech*, vol. 2013, pp. 2345-2349. 2013.
- Povey, Daniel, et al. "Purely sequence-trained neural networks for ASR based on lattice-free MMI." *Interspeech*. 2016.
- Hadian, Hossein, et al. "End-to-end Speech Recognition Using Lattice-free MMI." *Interspeech*. 2018.
- Kanda, Naoyuki, Yusuke Fujita, and Kenji Nagamatsu. "Lattice-free State-level Minimum Bayes Risk Training of Acoustic Models." *Interspeech*. 2018.

参考文献

- Manohar, Vimal, Daniel Povey, and Sanjeev Khudanpur. "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning." *2017 IEEE ASRU*. IEEE, 2017.
- Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.
- Ko, Tom, et al. "Audio augmentation for speech recognition." *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
- Snyder, David, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus." *arXiv preprint arXiv:1510.08484* (2015).
- Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).
- Park, Daniel S., et al. "Specaugment on large scale datasets." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- Medennikov, Ivan, Yuri Y. Khokhlov, Aleksei Romanenko, Dmitry Popov, Natalia A. Tomashenko, Ivan Sorokin, and Alexander Zatvornitskiy. “An Investigation of Mixup Training Strategies for Acoustic Models in ASR.”*Interspeech*, 2018.
- Ghahremani, Pegah, et al. "Investigation of transfer learning for ASR using LF-MMI trained neural networks." *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017.
- Manohar, Vimal, Daniel Povey, and Sanjeev Khudanpur. "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning." *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017.

参考文献

- Manohar, Vimal, et al. "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models." *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.
- Rousseau, Anthony. "Xenc: An open-source tool for data selection in natural language processing." *The Prague Bulletin of Mathematical Linguistics* 100.1 (2013): 73-82.
- Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. 2014.
- Liu, Qian, et al. "A General Procedure for Improving Language Models in Low-Resource Speech Recognition." *2019 International Conference on Asian Language Processing (IALP)*. IEEE, 2019.
- Liu, Xunying, et al. "Efficient lattice rescoring using recurrent neural network language models." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.
- Xu, Hainan, et al. "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.
- Li, Ke, et al. "An Empirical Study of Transformer-Based Neural Language Model Adaptation." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- Fiscus, Jonathan G. "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)." *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, 1997.
- Xu, Haihua, et al. "An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination." *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.
- Xu, Haihua, et al. "Minimum bayes risk decoding and system combination based on a recursion for edit distance." *Computer Speech & Language* 25.4 (2011): 802-828.

参考文献

- Tilk, Ottokar, and Tanel Alumäe. "LSTM for punctuation restoration in speech transcripts." Sixteenth annual conference of the international speech communication association. 2015.
- Tilk, Ottokar, and Tanel Alumäe. "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration." *Interspeech*. 2016.
- Xu, Kaituo, Lei Xie, and Kaisheng Yao. "Investigating LSTM for punctuation prediction." *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2016.
- Yu, Junjie, and Zhenghua Li. "Chinese spelling error detection and correction based on language model, pronunciation, and shape." *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 2014.
- Huang, Qiang, et al. "Chinese spelling check system based on tri-gram model." *Proceedings of the third CIPS-SIGHAN joint conference on Chinese language processing*. 2014.
- Guo, Jinxi, Tara N. Sainath, and Ron J. Weiss. "A spelling correction model for end-to-end speech recognition." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- Zhang, Shiliang, Ming Lei, and Zhijie Yan. "Automatic Spelling Correction with Transformer for CTC-based End-to-End Speech Recognition." *arXiv preprint arXiv:1904.10045* (2019).