语音关键词最新技术架构和在腾讯的应用实践

姓名：袁有根
公司：腾讯
时间：2021年10月16日

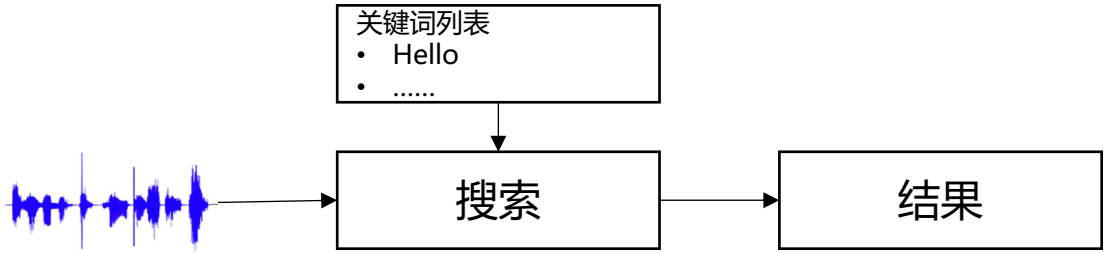
目录

目录

背景简介

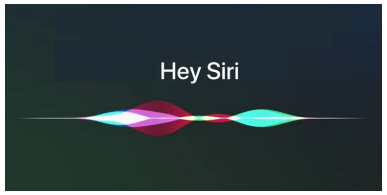
语音关键词检测技术目标

- 检测预定义的关键词



语音关键词检测应用场景

- 语音设备控制



- 海量语音文档快速检索

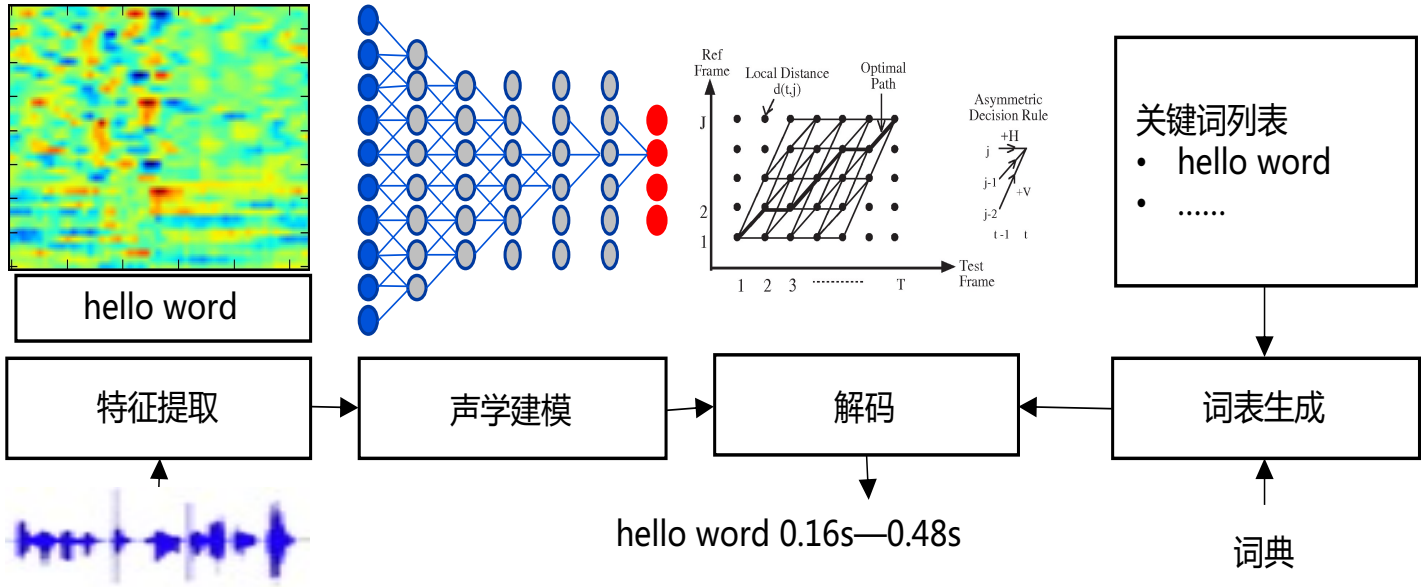
语音关键词检测主流方法

	KWS	WUW	QBE
技术路线	以文搜音	命令词唤醒	以音搜音
主流方法	基于语音识别	填充模型，端到端模型	模板匹配
优势	针对大量关键词，有时间位置信息，能重复检索	针对少数预定义关键词，检测速度极快	针对部分关键词，常用语低资源场景
劣势	速度慢	灵活性不强	效果差，依赖模板质量

背景简介

语音关键词检测基线系统

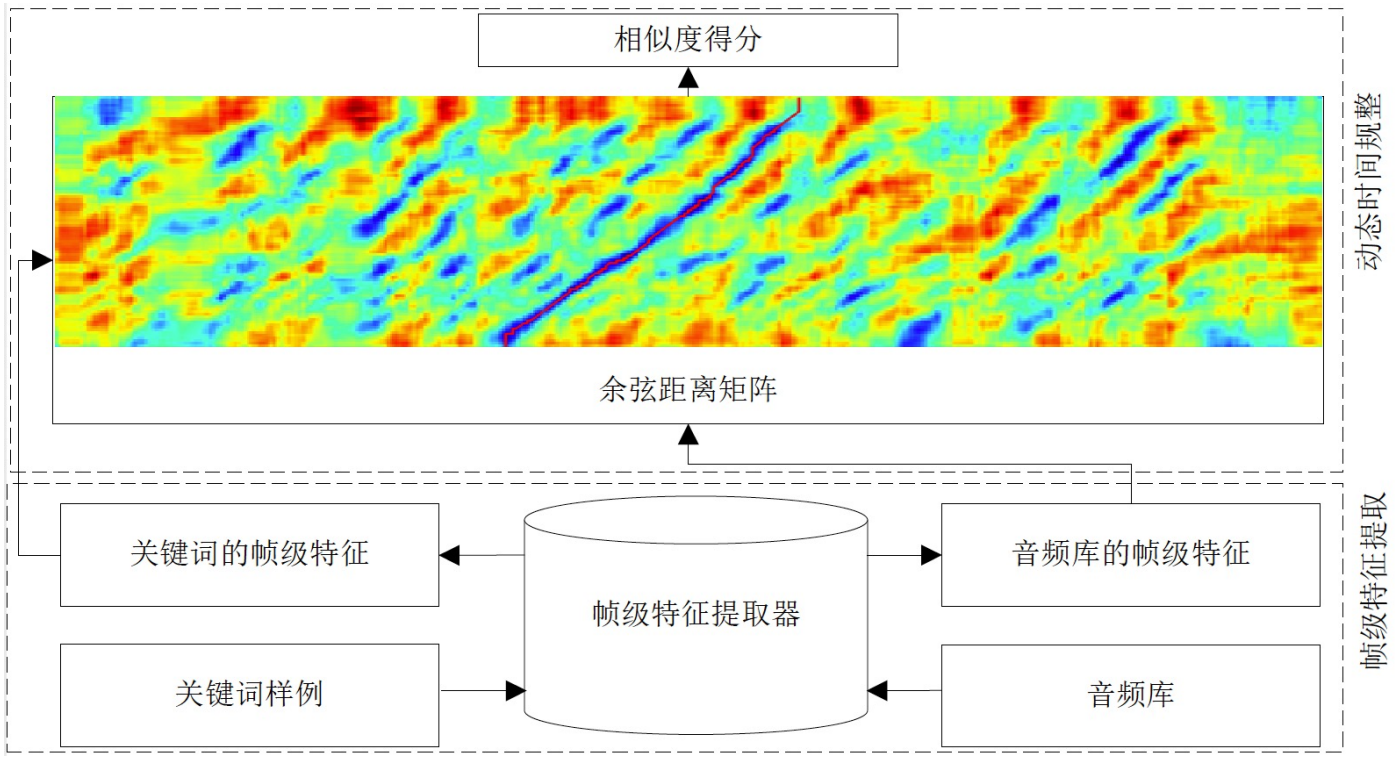
- 定点化特征提取
  - ✓ 比Kaldi快1倍
- tdnn声学模型
  - ✓ 比DNN效果更好
  - ✓ 三倍跳帧，推理速度更快
- 自定义词典
- 基于WFST的解码器
  - ✓ 无语言模型重打分，搜索速度更快
  - ✓ 得分归一化



目录

基于帧级声学特征的QBE

- 帧级声学特征提取
  - ✓ posteriorgram features [1]
  - ✓ Autoencoder features [2]
  - ✓ Bottleneck features [2]
- 声学模板匹配
  - ✓ DTW

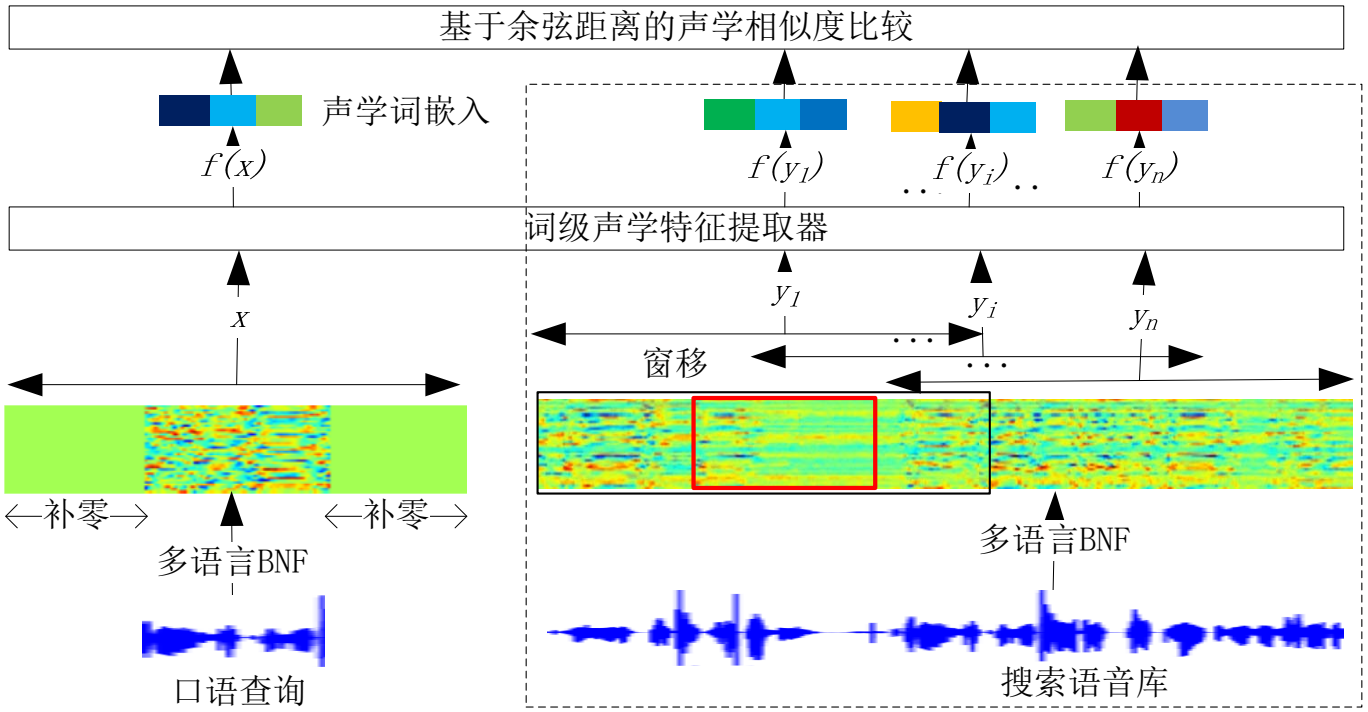


[1] Hazen T J, Query-by-example spoken term detection using phonetic posteriorgram templates[C]. ASRU. 2009: 421-426.

[2] Yougen Yuan, et al. Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection [C]. ICASSP. 2017: 56455-55649

基于词级声学特征的QBE

- 词级声学特征提取
  - ✓ Siamese CNN [1]
  - ✓ Tripelt RNN [2]
  - ✓ Multi-view BLSTM [2]
- 声学模板匹配
  - ✓ DTW



[1] Kamper H. Deep convolutional acoustic word embeddings using word-pair side information[C]. ICASSP, 2016: 4950-4954.

[2] Settle S. Discriminative acoustic word embeddings: Tecurrent neural network-based approaches[C]. SLT, 2016: 503-510.

[3] He W. Multi-view recurrent neural acoustic word embeddings[J]. arXiv preprint arXiv:1611.04496, 2016.

技术对比

	基于帧级声学特征的QBE	基于词级声学特征的QBE
优势	有/无监督场景都适用	速度快，整词信息
劣势	速度慢，效果差	依赖词语边界

应用思考

- QBE方法主要适用于低资源场景，效果较差
- QBE技术很难直接应用，要想落地必须结合特定的任务

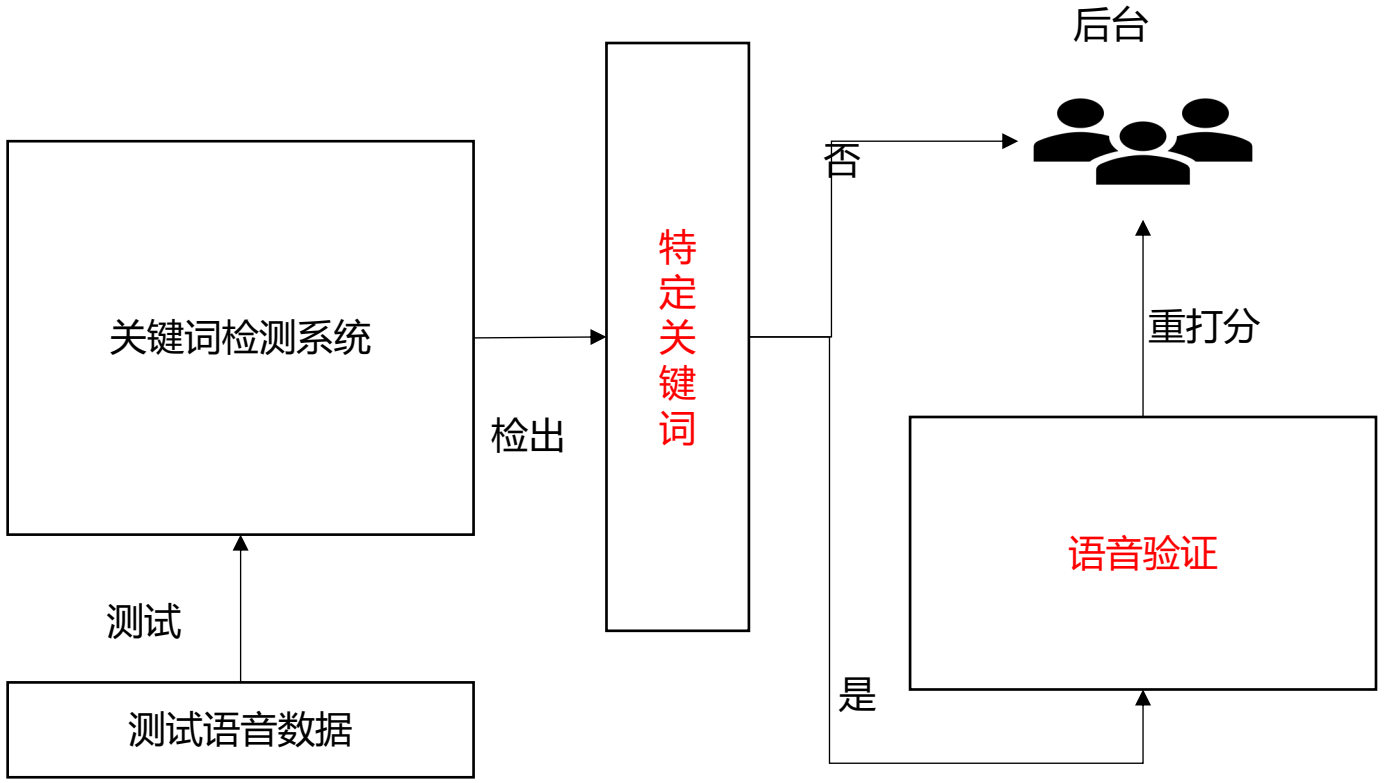
QBE技术与应用

关键词检测技术短板

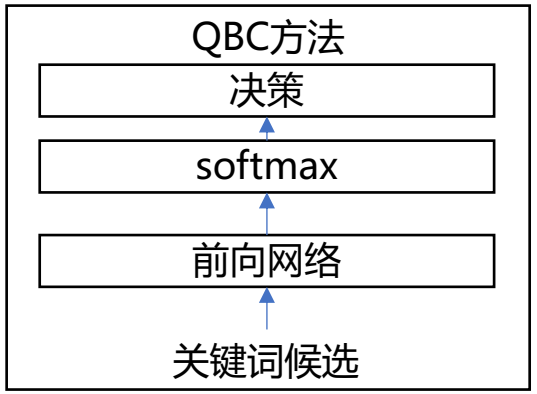
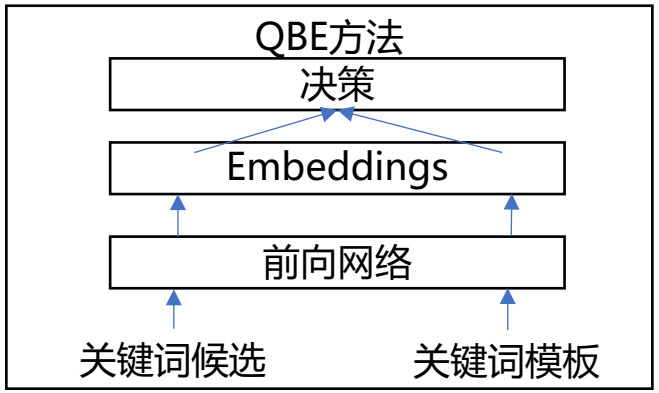
- 基线系统无法对特定关键词进行优化

方案制定

- 增加关键词黑名单机制
- 增加语音验证模块



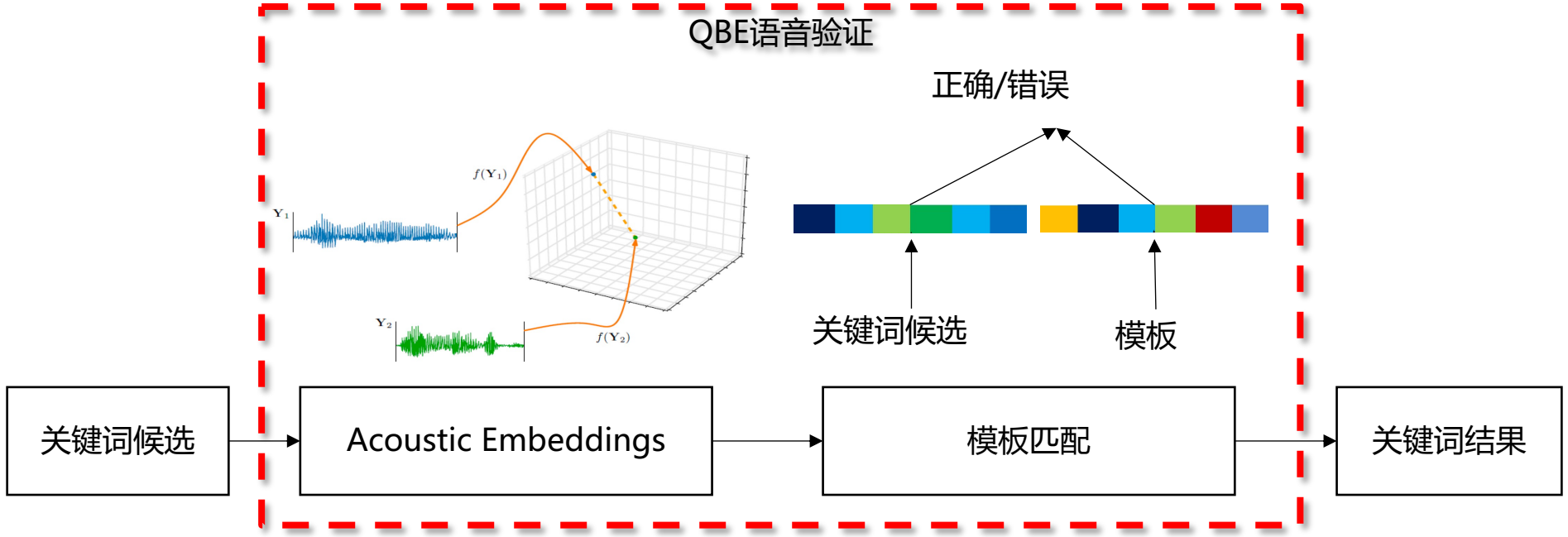
技术对比

	实例分类(QBC)	实例查询(QBE)
定义	直接输出类别标签	用模板进行判定，“去伪存真”
方法		
优势	简单直接	效果更好，设置灵活
劣势	关键词变更都要重新训练	依赖模板质量

QBE语音验证 vs 语音关键词检测基线

- 充分利用真实场景的Case，“变废为宝”
- 整词建模，对词语的区分能力更强 [1]

系统流程



[1] Yougen Yuan, et al. Learning acoustic word embeddings with temporal context for query-by-example speech search[C]. INTERSPEECH. 2018: 97-101.

Acoustic Embeddings 学习

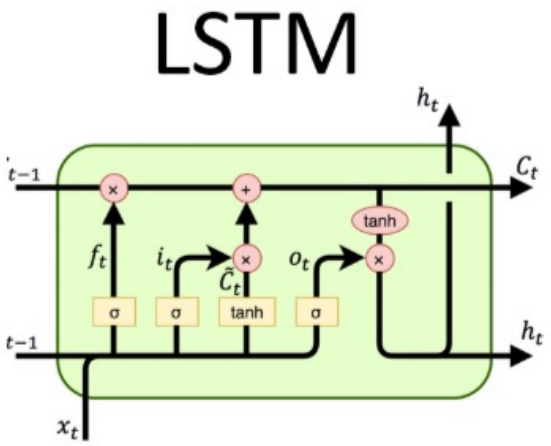
- 在Triplet中将虚警case作为负样例

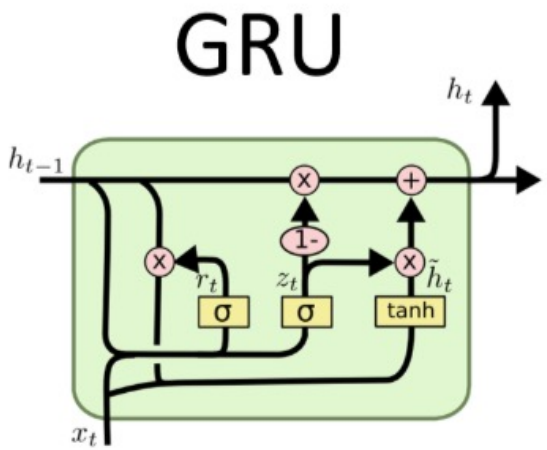
负样例选择

非关键词case

关键词虚警case

- Bidirectional LSTM → Bidirectional GRU

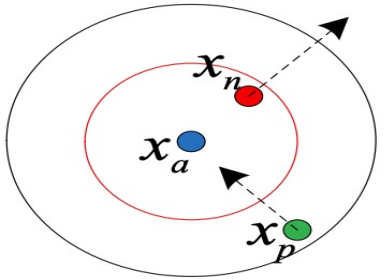




Acoustic Embeddings 学习

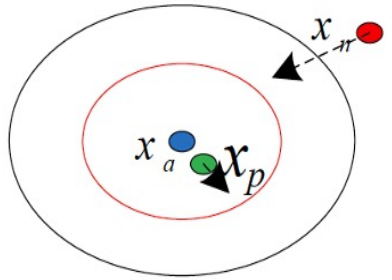
- 三体损失

$$TL(x_a, x_p, x_n) = \max\{0, d_+ - d_- + \delta_1\}$$



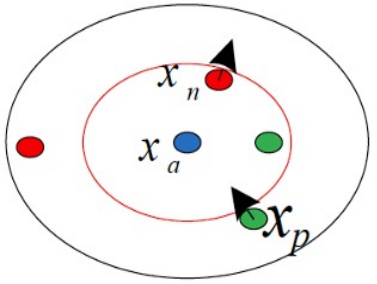
- 反向三体损失

$$RTL(x_a, x_p, x_n) = \max\{0, d_- - d_+ - \delta_2\}$$



- 铰链损失

$$HL(x_a, x_p, x_n) = \max\{0, -\theta + d_+\} + \max\{0, \theta - d_-\}$$



模板匹配

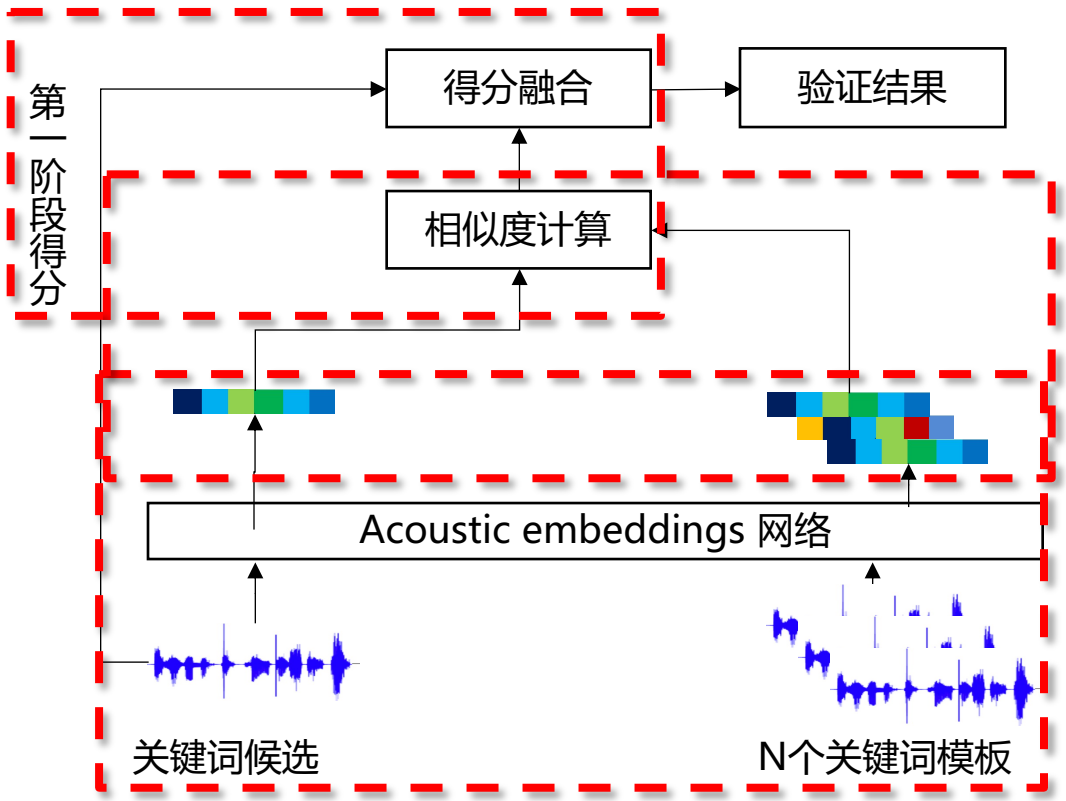
- Embeddings 提取
  - ✓ 模板Embeddings预提取
  - ✓ 特征复用

- 相似度计算

$$\text{Similary}(x, y) = \frac{1 - \cos(f(x), f(y))}{2}$$

- 得分融合

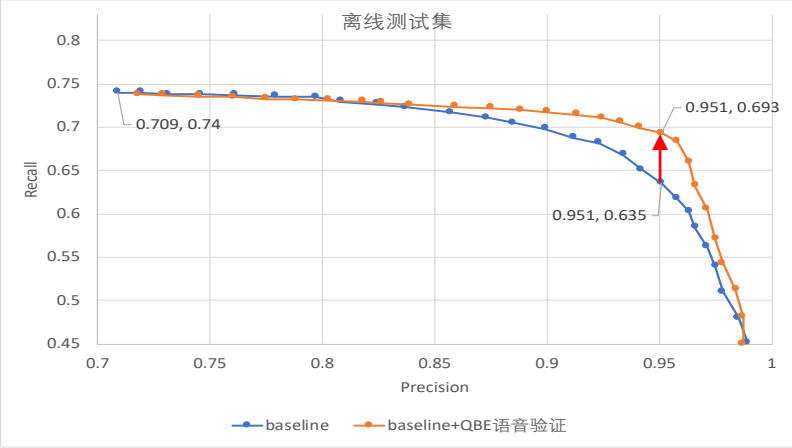
$$\text{Score} = (1 - a)Score_{stage1} + (a)Score_{stage2}$$



QBE语音验证

离线效果

- 在离线测试集[1]中，QBE语音验证中每个优化方法在相同准确率下的覆盖都有一点提升
- 相比baseline，最终的QBE语音验证模型在相同准确率下的Recall从0.635提升到0.693



方法	Precision	Recall
QBE语音验证	0.95	0.60
+虚警做负样例	0.95	0.62
+BGRU	0.95	0.64
+ 损失函数改进	0.95	0.67
+ 得分融合	0.95	0.69

工作小结

- 项目背景：特定关键词漏过较多
- 技术方法：QBE语音验证
- 创新点：更强的embddings模型+更快的模板匹配

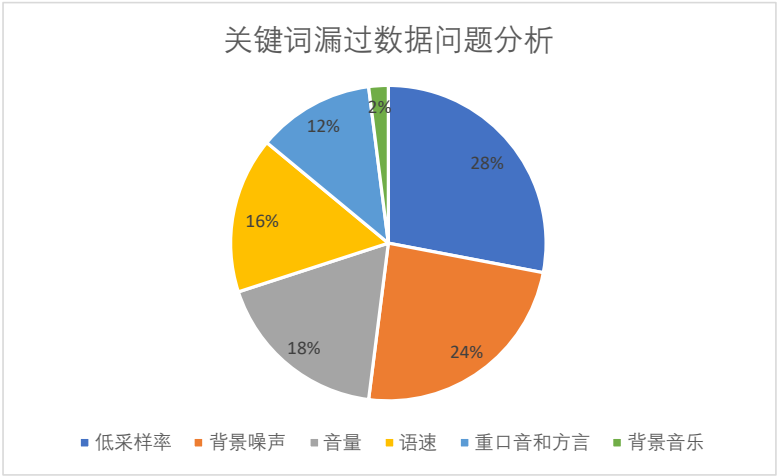
[1] 离线测试集主要是一个内部标注的技术短板测试集

目录

Hybrid语音关键词检测

项目背景

- 复杂声学场景下的关键词漏过偏多

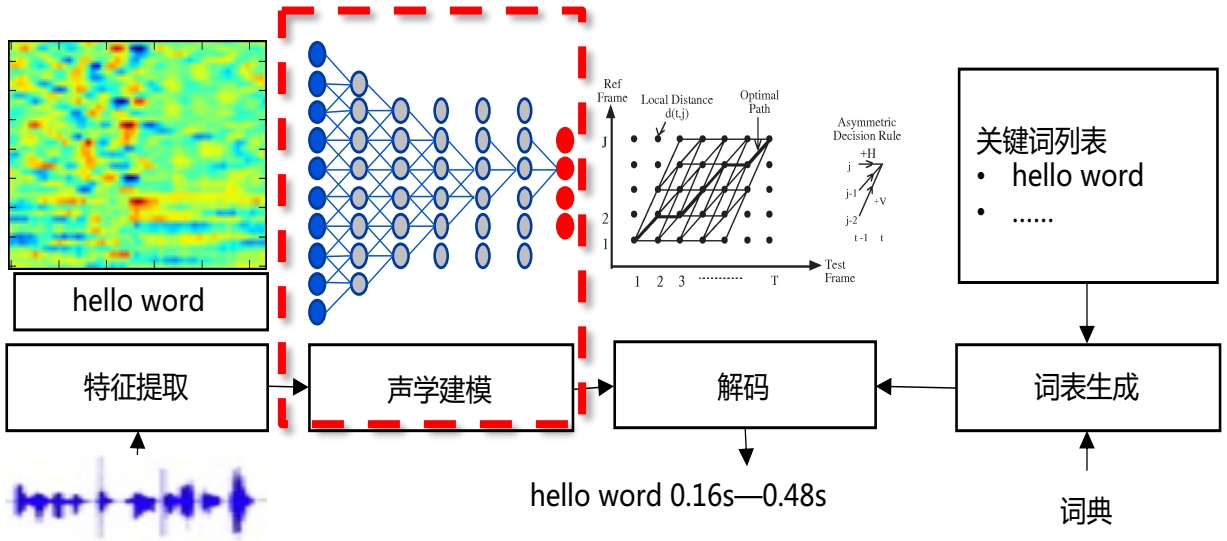


技术现状

- 基线系统的抗干扰能力较差

方案制定

- 增大建模单元的颗粒度
- 探索更强的声学模型
- 降低模型的计算复杂度

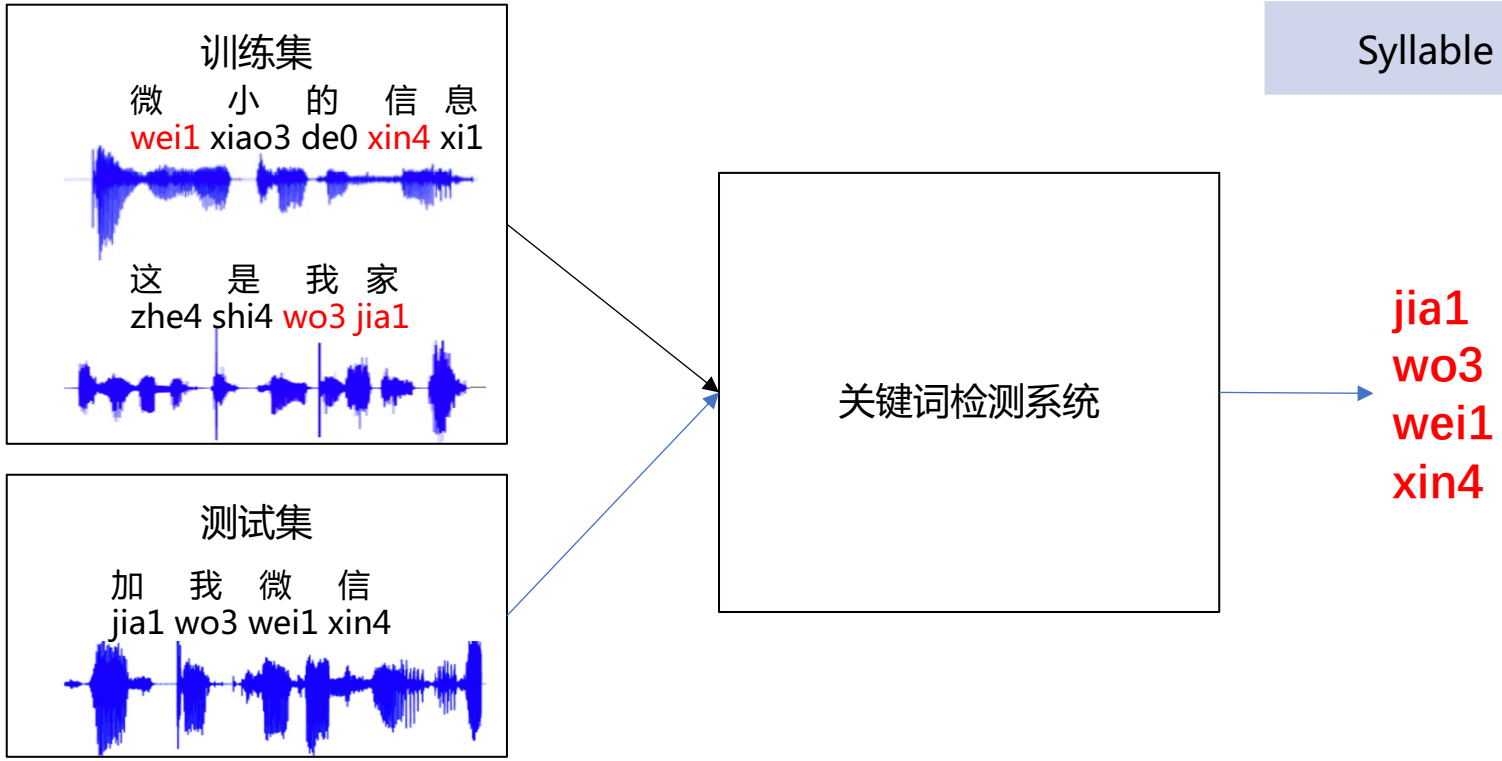


Hybrid语音关键词检测

方案制定

- 使用Syllable替代Phone作为建模单元
  - ✓ 上下文信息更多，对同音字也能恢复

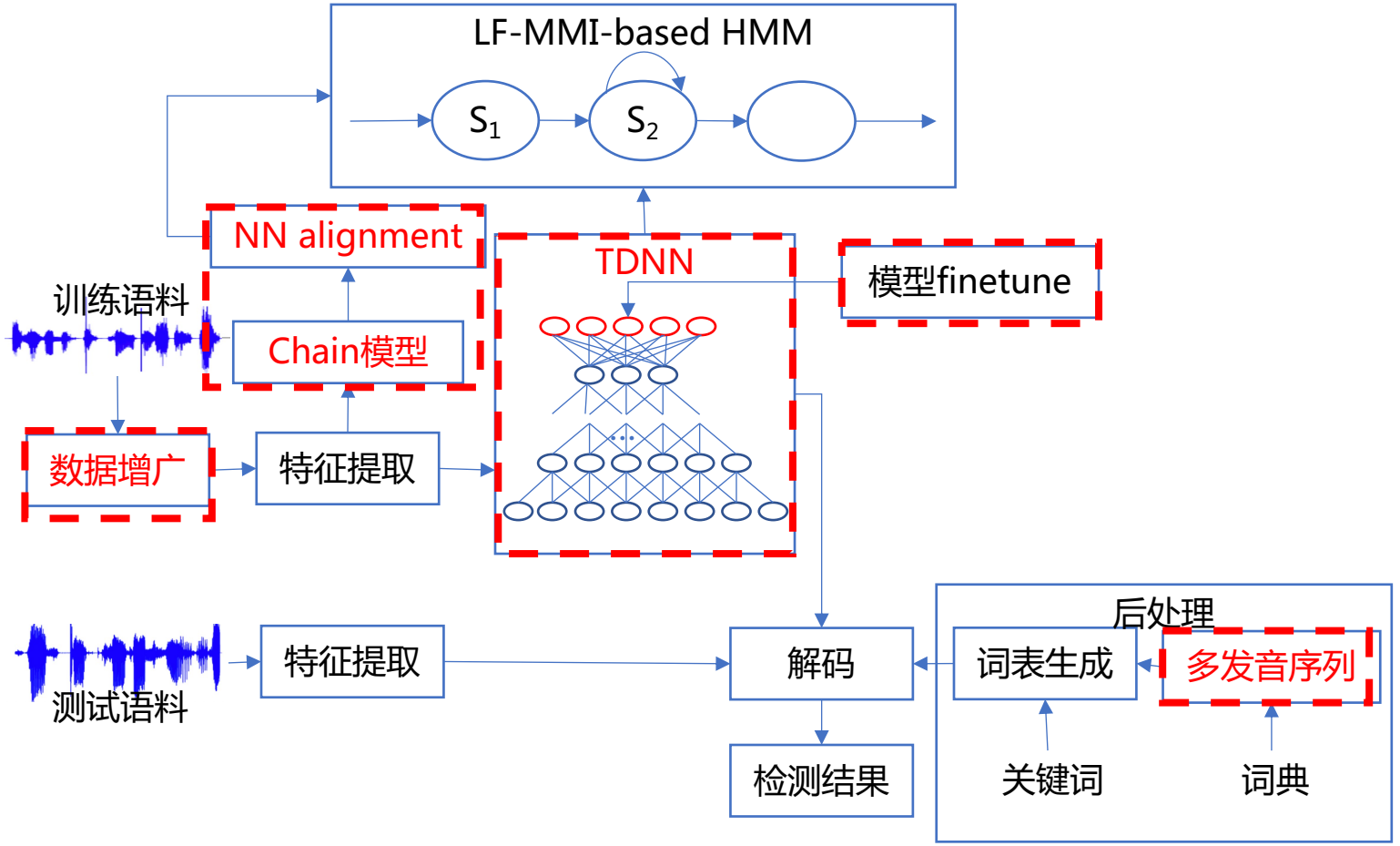
建模单元	关键词
Phone	g aa1 g aa1
Character	加 家
Syllable	jia1 jia1



Hybrid语音关键词检测

方案制定

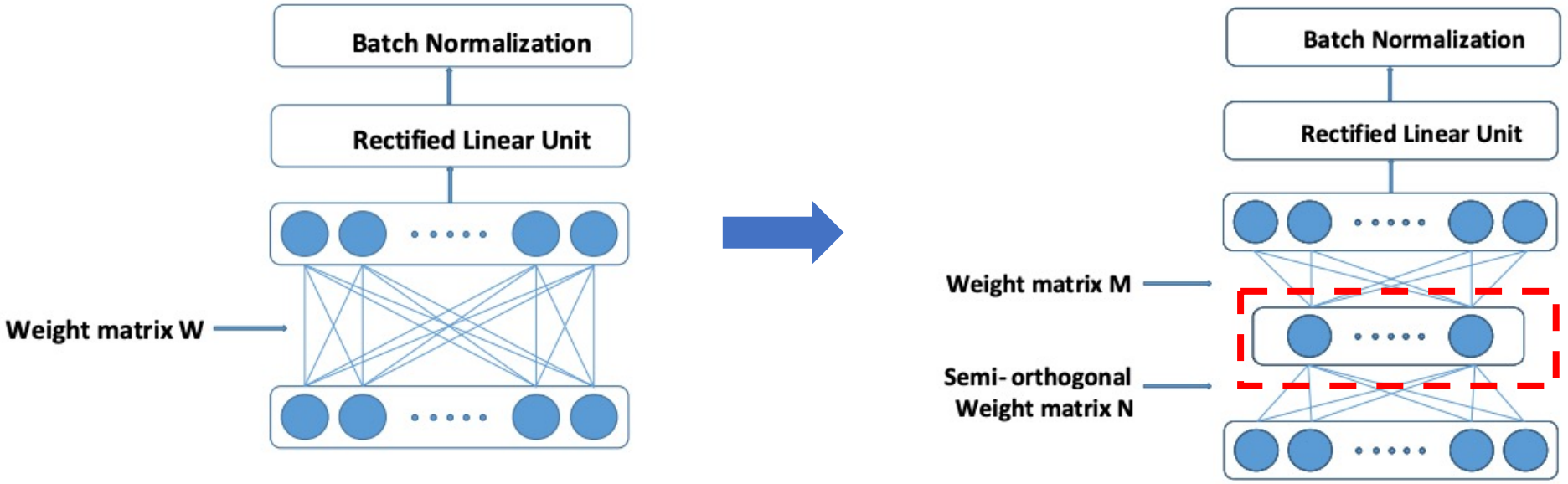
- 使用LF-MMI替代CE指导声学模型训练



Hybrid语音关键词检测

方案制定

- 在TDNN网络中插入SVD[1]加速推理

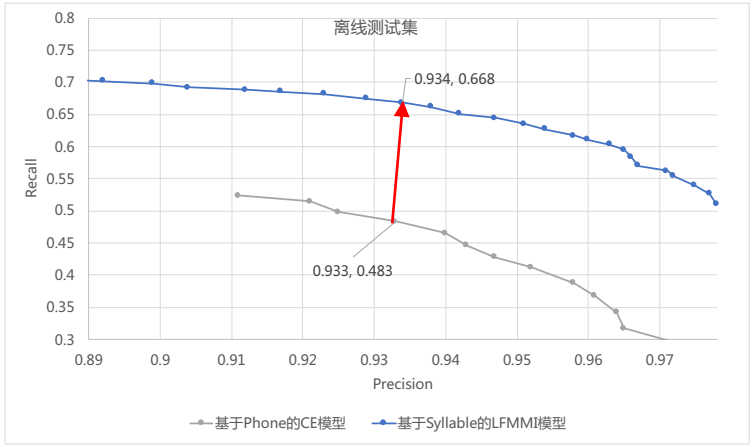


[1] Povey D. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks[C]//Interspeech. 2018: 3743-3747.

Hybrid语音关键词检测

离线效果

- 每个优化方法在相同准确率下的覆盖都有提升
- 相比baseline，基于音节的LFMMI模型在相同准确率下的Recall从0.483提升到0.668



方法	Precision	Recall
Baseline	0.934	0.535
+ 数据增广	0.934	0.585
+ NN帧对齐	0.934	0.608
+ 多发音序列	0.934	0.646
+改用TDNN-F	0.934	0.651
+模型finetune	0.934	0.668

工作小结

- 项目背景：复杂场景下关键词漏过较多
- 技术方法：基于Syllable的LFMMI模型
- 创新点：Syllable建模（提升覆盖）+LFMMI（提升准确）+TDNN-F（加速推理）

目录

End2end语音关键词检测

项目背景

- 关键词检测的覆盖依然不够高

技术现状

- Hybrid模型的性能接近天花板，优化提升非常有限
- 端到端模型在学术研究中不断取得突破

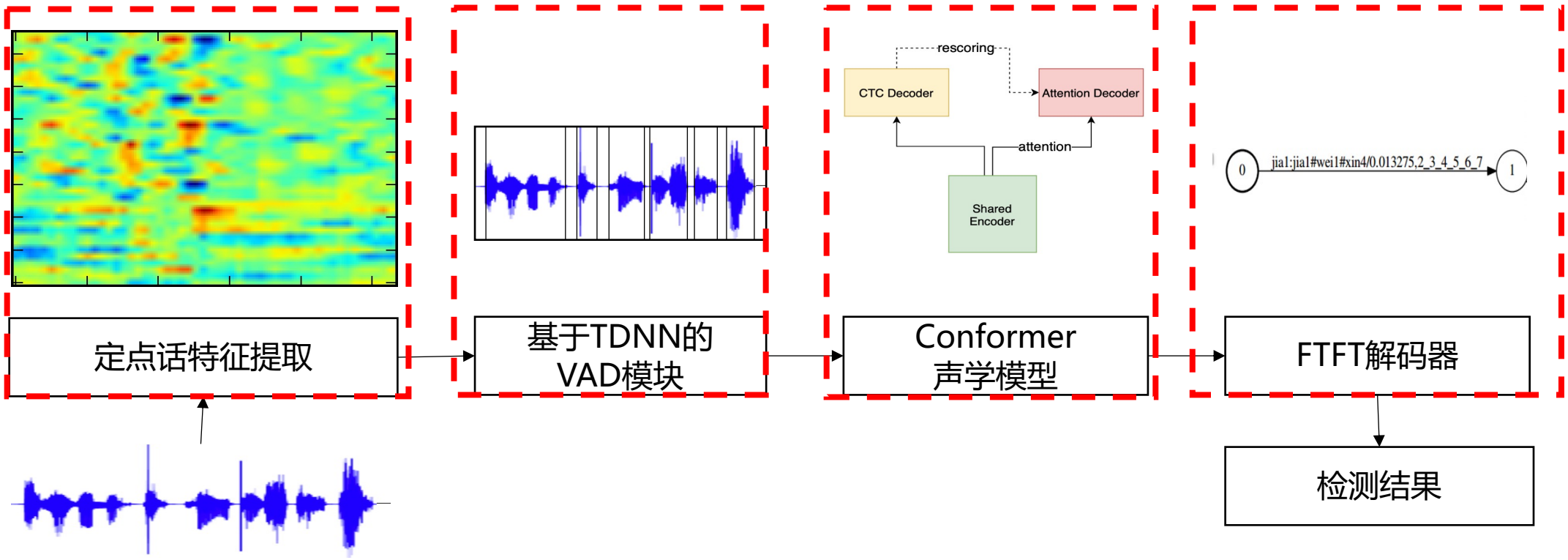
方案制定

- 使用端到端模型替代Hybrid模型

语音关键词检测	Hybrid模型	端到端模型
方法	需要声学、词典和模型	用单个模型直接从特征到文本序列
优点	鲁棒性强	效果好
缺点	效果有局限、优化空间少	灵活性较差

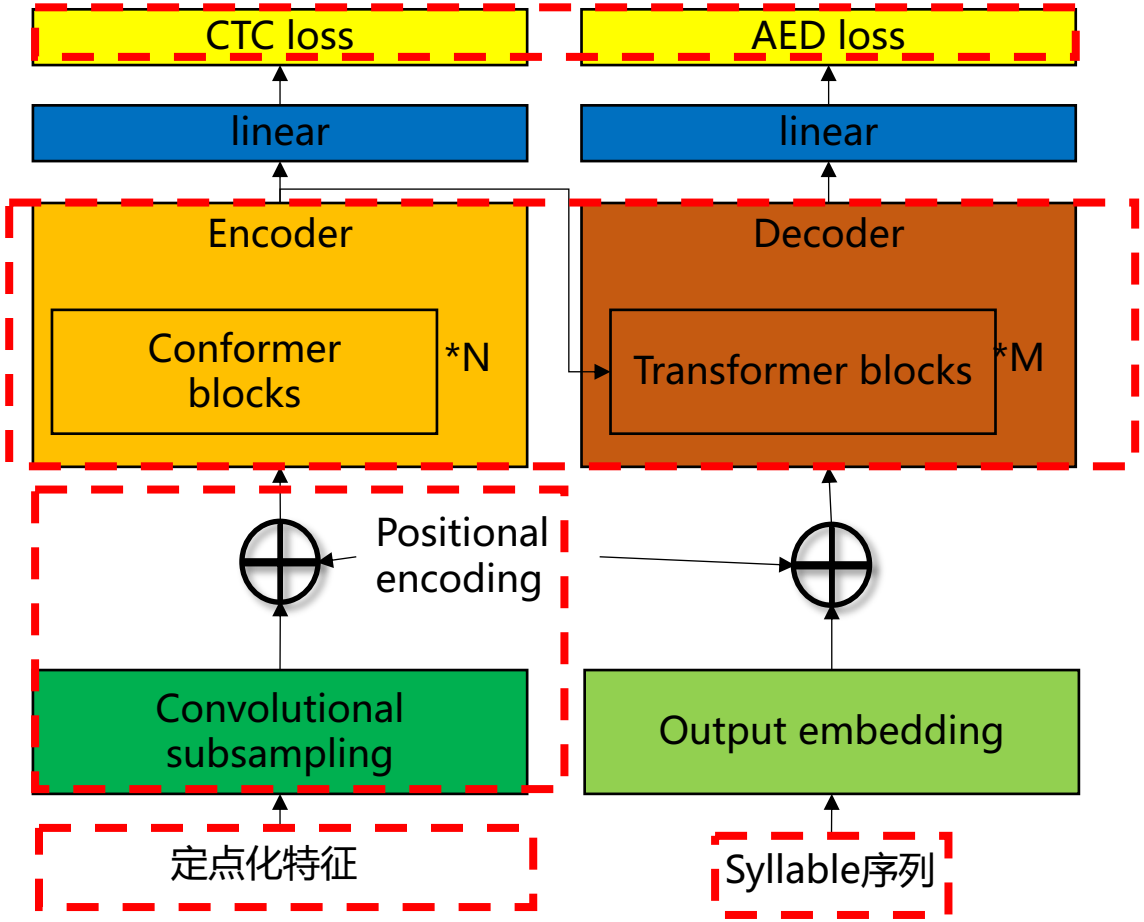
End2end语音关键词检测

系统流程



End2end语音关键词检测

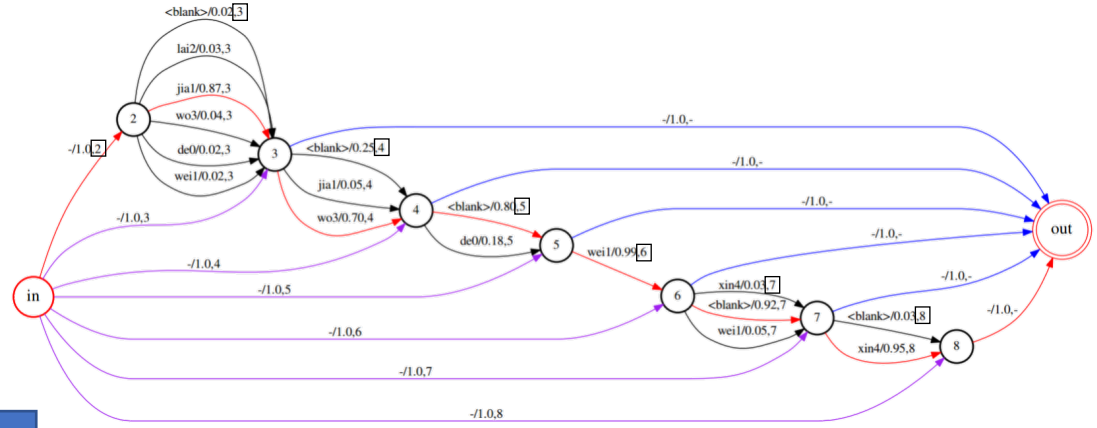
Conformer声学模型



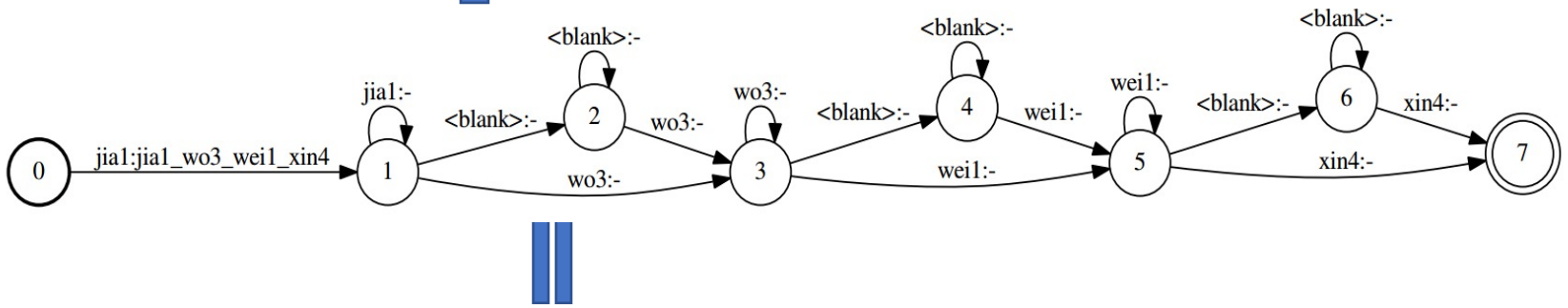
End2end语音关键词检测

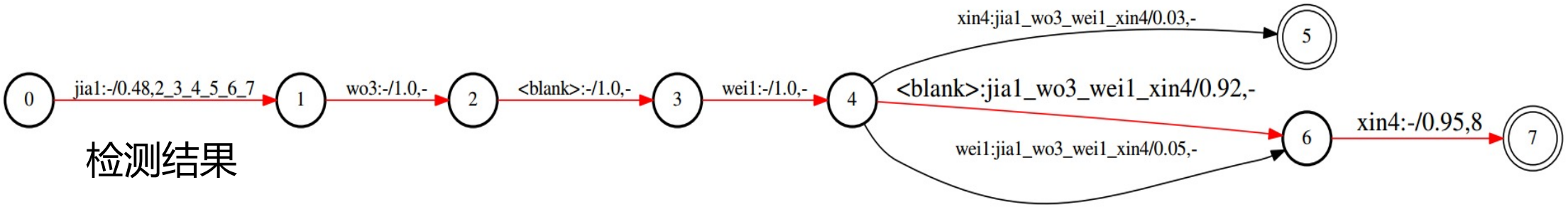
FTFT解码

Fast Timed Factor Transducer索引图



关键词词表解码图



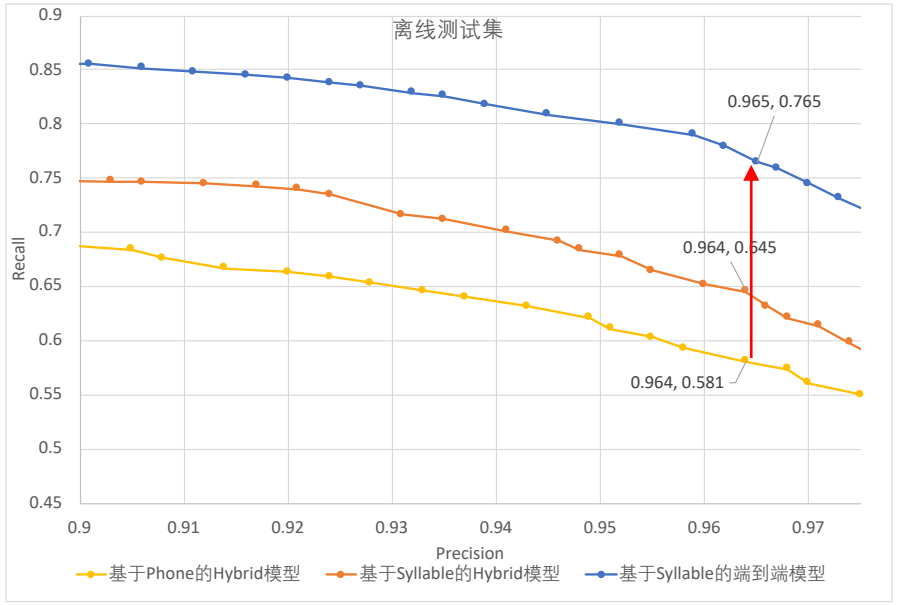


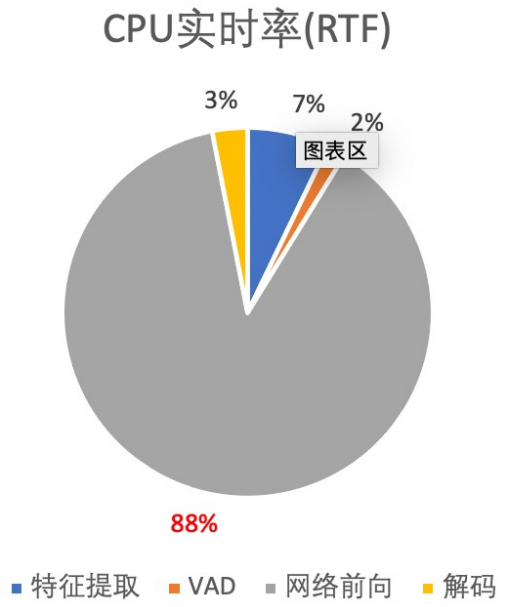
检测结果

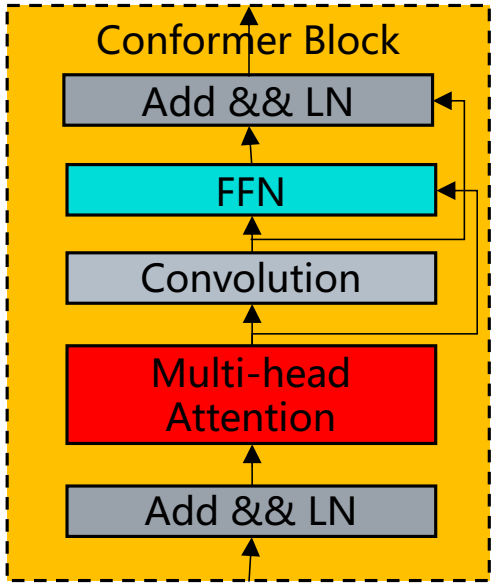
End2end语音关键词检测

离线效果

- 相比baseline，基于Syllable的端到端模型在相同准确率下的Recall从0.581提升到0.765







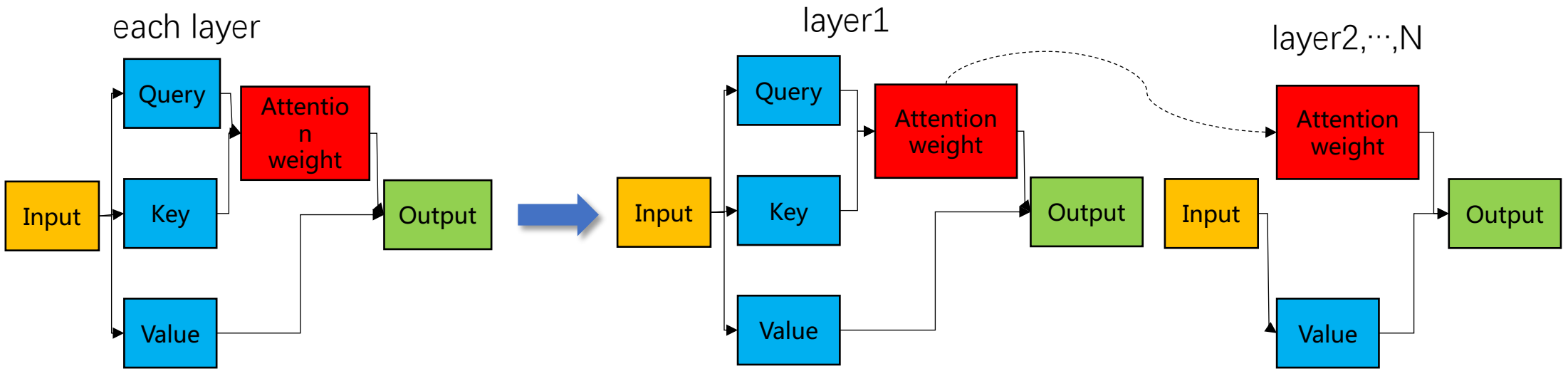
技术短板

- 端到端模型计算量较大，88%的推理时间用在Conformer Block网络前向计算上

End2end语音关键词检测

推理加速

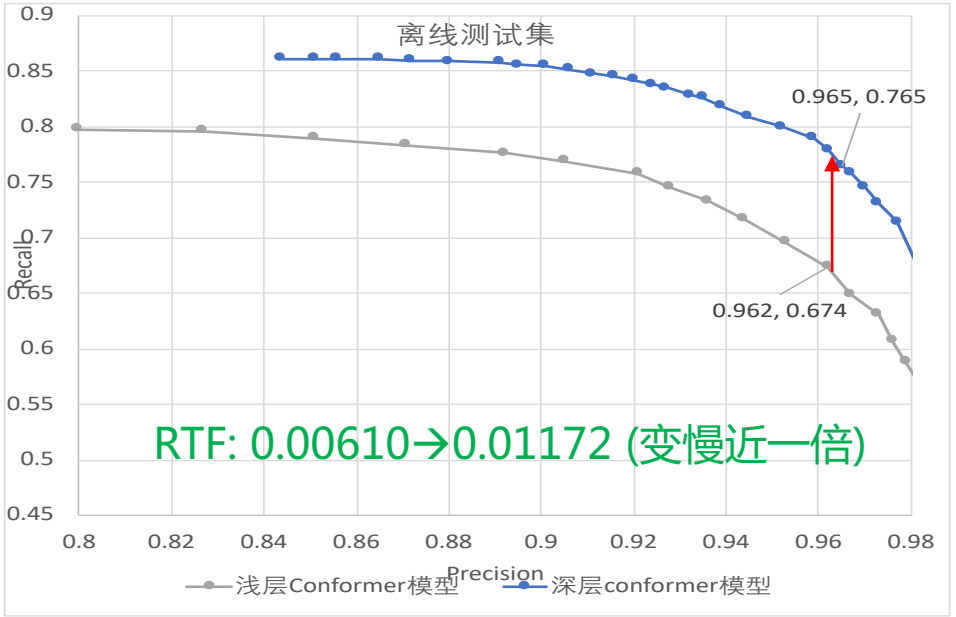
- 问题：Attention机制导致conformer模型的计算量增大
- 分析：Query和Key就是用来计算注意力权重的，每一层的权重值变化很小
- 方法：注意力权重复用

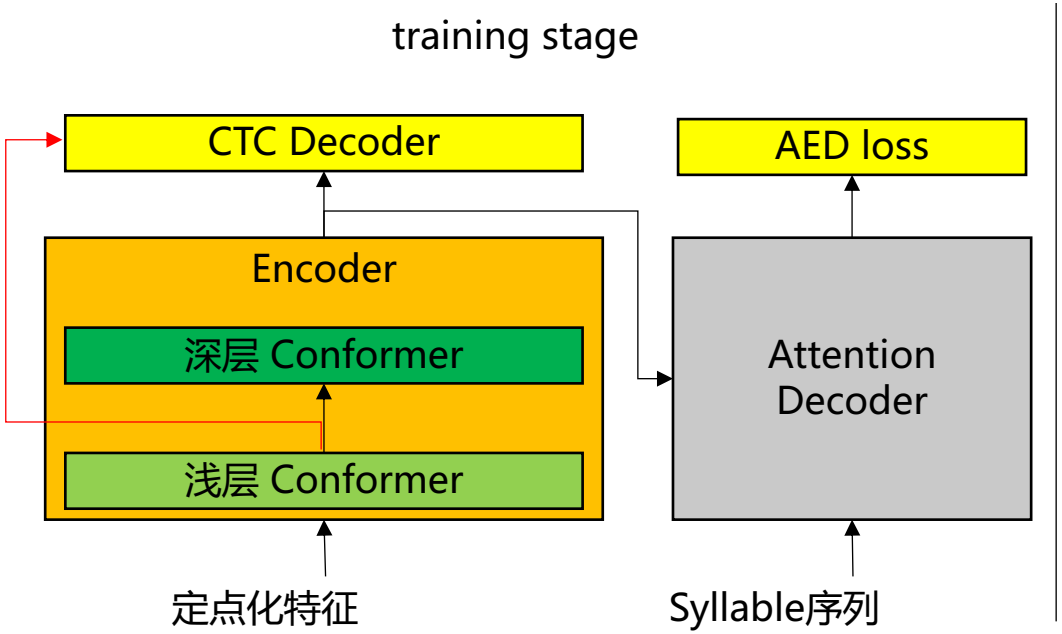


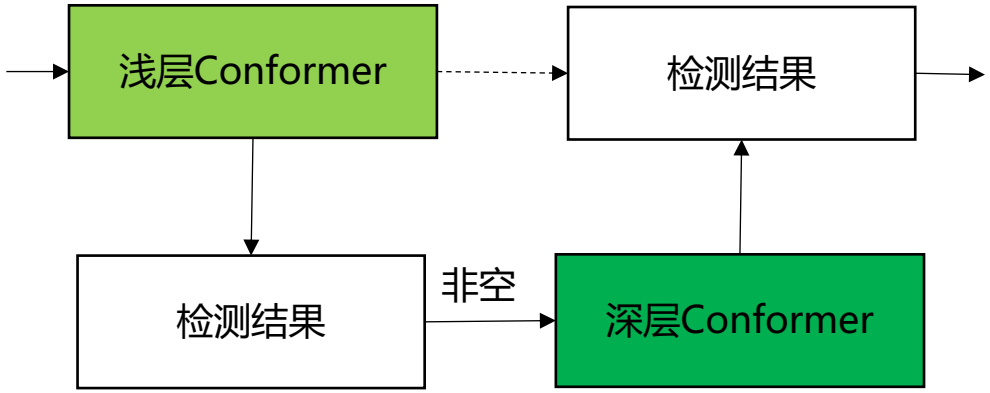
End2end语音关键词检测

推理加速

- 问题：深层模型效果比浅层好，但是推理速度明显变慢
- 思考：寻找一个方法既能保证效果，又能加速推理
- 方法：构建层次网络，浅层保覆盖，深层保准确





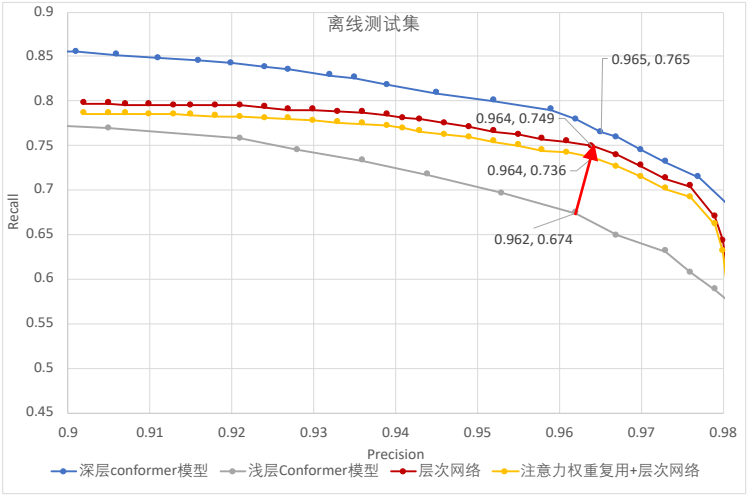


注：浅层模型主要指3层的conformer模型；深层模型主要指12层的conformer模型

End2end语音关键词检测

离线效果

- 相比浅层模型，层次网络在相同准确率下Recall从0.674提升到0.749，但是推理速度变化不大
- 加上SVD和注意力权重后，CPU实时率还能加快17%



系统	CPU实时率
深层Conformer模型	0.01172
浅层Conformer模型	0.00610
层次网络	0.00625
注意力权重复用 + 层次网络	0.00550(↑12%)

工作小结

- 项目背景：关键词检测覆盖依然不够高
- 技术方法：基于层次Conformer的end2end模型
- 创新点：系统升级+推理加速

目录

总结

QBE技术与应用

- 在声学特征层面，探索了更好的QBE方法
- 改进了QBE用作语音验证，使得特定关键词的漏过变少

Hybrid语音关键词检测

- 在声学建模层面，全面提升系统鲁棒性
- 使用了基于Syllable的LFMMI模型，使得复杂场景下的关键词漏过变少

End2end语音关键词检测

- 优化了关键词检测的系统架构
- 使用了基于层次Conformer的end2end模型，使得系统效率明显提升

总结

- **VKW Hybrid关键词检测系统已经开源**
  - https://github.com/VKW2021/kaldi-baseline
- **VKW End2end关键词检测系统正在开源到wenet上**
  - pull request： https://github.com/wenet-e2e/wenet/pull/676
- **希望更多的技术开源工作，共同促进技术发展**

展望

- 更高效的语音关键词检测系统
  - 系统性能层面
  - 计算资源，推理速度层面
- 更灵活的技术框架和路线
  - 声学建模层面
  - 系统模块化
- 更广阔的技术落地场景
  - 多语种
  - 流式

Thanks