中国计算机学会 语音对话与听觉专委会

【语音之家】走进企业系列沙龙

语音技术在 58同城 的应用

2021/8/31 18:30-20:30

沙龙形式：线上直播

简介： 近年来，AI智能语音应用在58同城广泛落地，从2018年开始，58同城AI Lab先后研发了语音对话机器人、语音内容分析平台，在销售、客服、产品、运营场景广泛使用，打造了多款AI应用，如销售智能外呼助手、呼叫中心语音质检系统、招聘面试机器人等。其中，语音对话机器人、语音内容分析平台底层依赖语音识别技术，初期由第三方语音厂商提供服务，但在2020年7月，58自主研发的语音识别引擎正式上线，替换了第三方语音厂商。当前已接入了数十个不同业务场景，日均转写语音量达数万小时。

本次沙龙将分享58同城AI Lab自研语音识别引擎的历程，包括数据、算法和工程架构，以及在58同城本地服务（黄页）业务中落地的一款智能语音应用——销售智能外呼助手。

沙龙议程

18:30 19:30	分享内容:58同城语音识别自研之路 李咏泽 | 58同城TEG-AI Lab算法资深工程师 王　焱 | 58同城TEG-AI Lab后端架构师
19:30 20:30	分享内容:电销场景下智能外呼语音机器人落地实践 李　忠 | 58同城TEG-AI Lab智能语音部负责人

主持人：詹坤林
58同城TEG AI Lab负责人
算法资深架构师，技术委员会AI分会主席

嘉宾：李咏泽
58同城TEG-AI Lab算法资深工程师
分享内容：58同城语音识别自研之路

嘉宾：王焱
58同城TEG-AI Lab后端架构师
分享内容：58同城语音识别自研之路

嘉宾：李忠
58同城TEG-AI Lab智能语音部负责人，算法高级架构师
分享内容：电销场景下智能外呼语音机器人落地实践

主办单位

CCF语音对话与听觉专委会
中国人工智能产业发展联盟（AIIA）评估组
58同城AI Lab
语音之家（北京）科技有限公司
北京希尔贝壳科技有限公司

合作社区

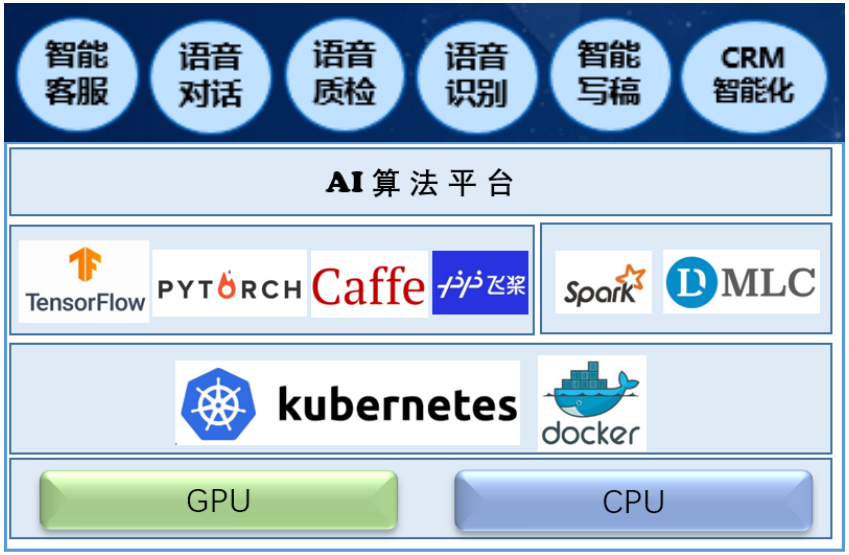
CSDN　Speechhome　segmentfault　掘金　OSCHINA

58AILab公众号

语音之家公众号

个人介绍 – 詹坤林

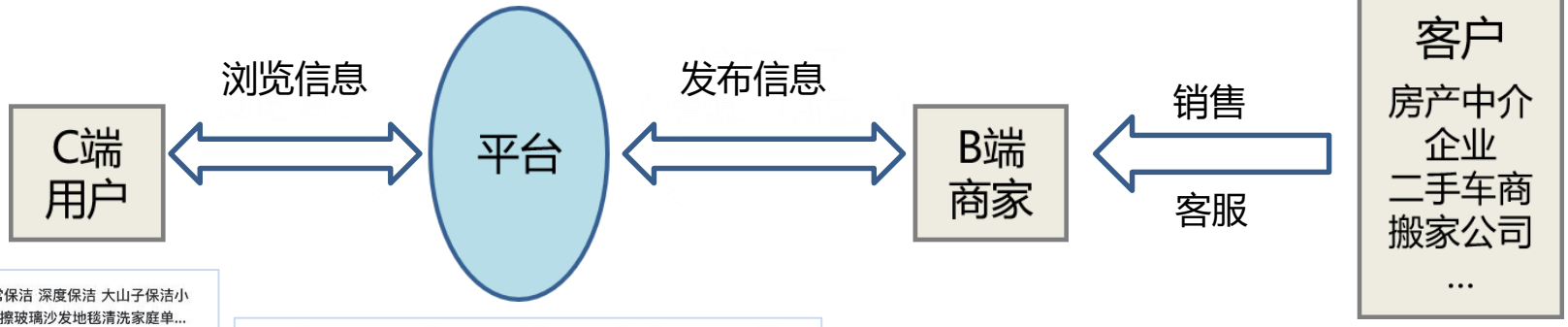
- 2015年5月加入58同城，现任TEG AI Lab负责人、算法资深架构师，兼任技术委员会 AI分会主席
- 加入58前任腾讯高级工程师，从事推荐算法研发
- 2012年硕士毕业于中国科学院大学，研究方向为数据挖掘



推动AI技术在58同城的落地，打造AI中台能力，以提升前台业务人效、收入和用户体验

58同城生活服务平台

让 生 活 简 单 美 好



语音场景：

电话语音：呼叫中心、C2B / B2C隐私通话

网络音频：微聊、短视频、直播

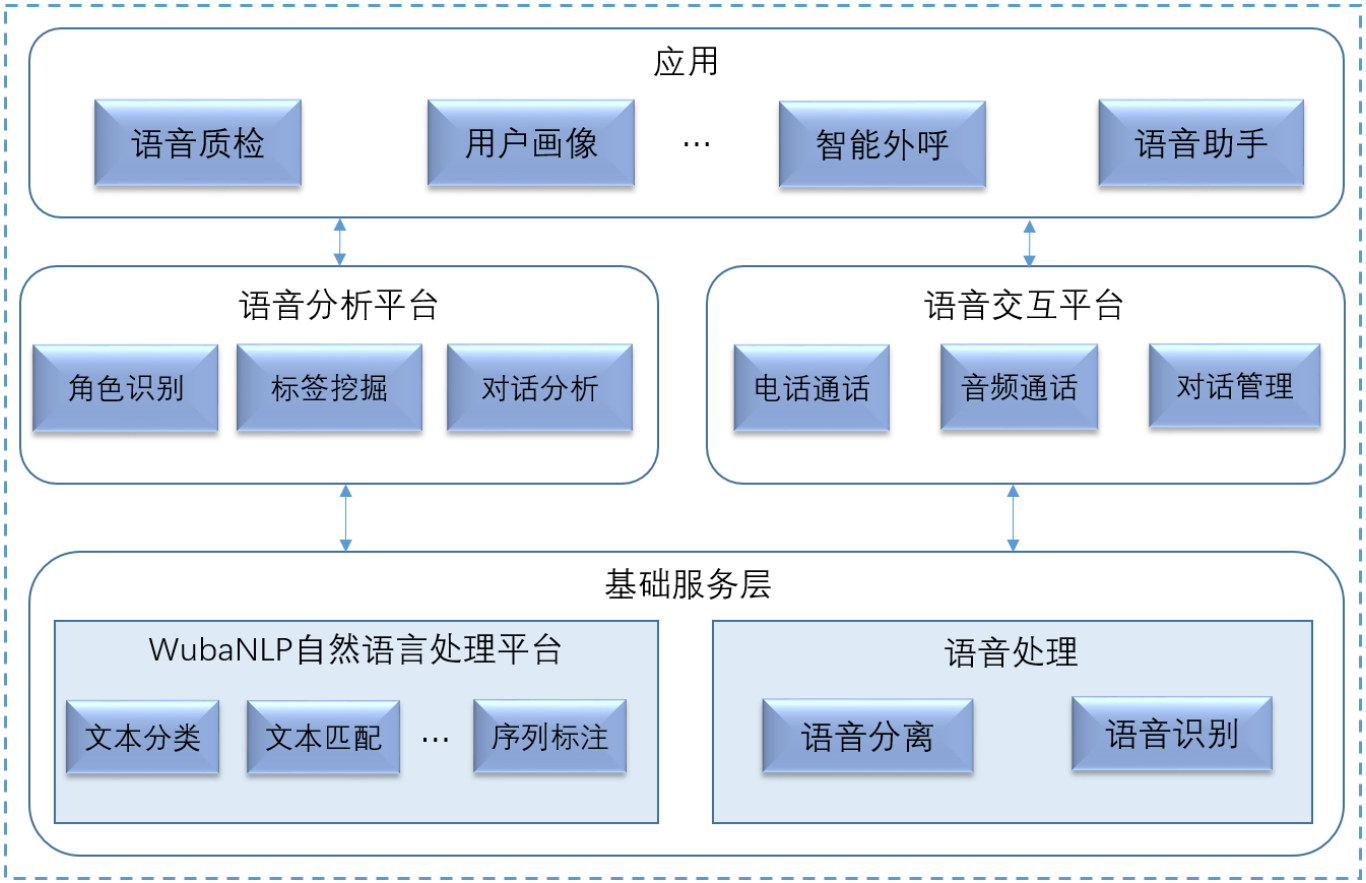
30+业务场景，语音量近千万小时/年

四大业务线：房产（安居客）、招聘、车、本地服务（黄页）

公司方向：全面拥抱**产业互联网**，在58主要领域里，努力成为全球第一家完成**产业化升级**的分类信息平台。

58同城语音语义中台

让 生 活 简 单 美 好



2018年10月上线语音机器人，2019年5月上线语音分析平台，2020年7月上线自研语音识别引擎。

58同城语音识别自研之路

语音算法　主讲人：李咏泽

工程架构　主讲人：王　焱

语音算法部分

李咏泽

个人介绍

- 工作经历
  - 2019年11月~至今 58同城 语音识别算法
  - 2018年10月~2019年10月 Rokid 声纹识别算法
- 教育背景
  - 2016年9月~2018年9月 硕士 巴黎第六大学
  - 2012年9月~2016年9月 本科 武汉大学

目录

智能语音在58的应用背景

58同城生活服务平台

智能语音在58的应用背景

让 生 活 简 单 美 好

呼叫中心销售客服质检

语音质检是什么

- 传统语音质检通常是指质检员听取一定比例的电话录音进行人工质检，检测坐席在通话过程中是否有违规或非标准话术的行为

58同城呼叫中心简介

- 支撑数千名销售、客服人员工作
- 年通话数 1亿+，通话时长数百万小时
- 采用传统方式开展质检
- 纯人工抽检录音，覆盖率低
- 人工质检效率低，人均单日可听3小时录音

智能语音在58的应用背景

让 生 活 简 单 美 好

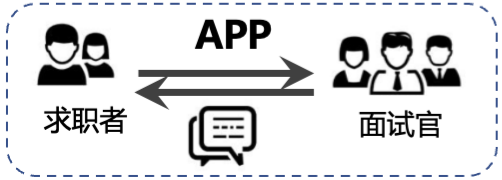
C端用户与B端商家隐私通话、微聊

智能语音在58的应用背景

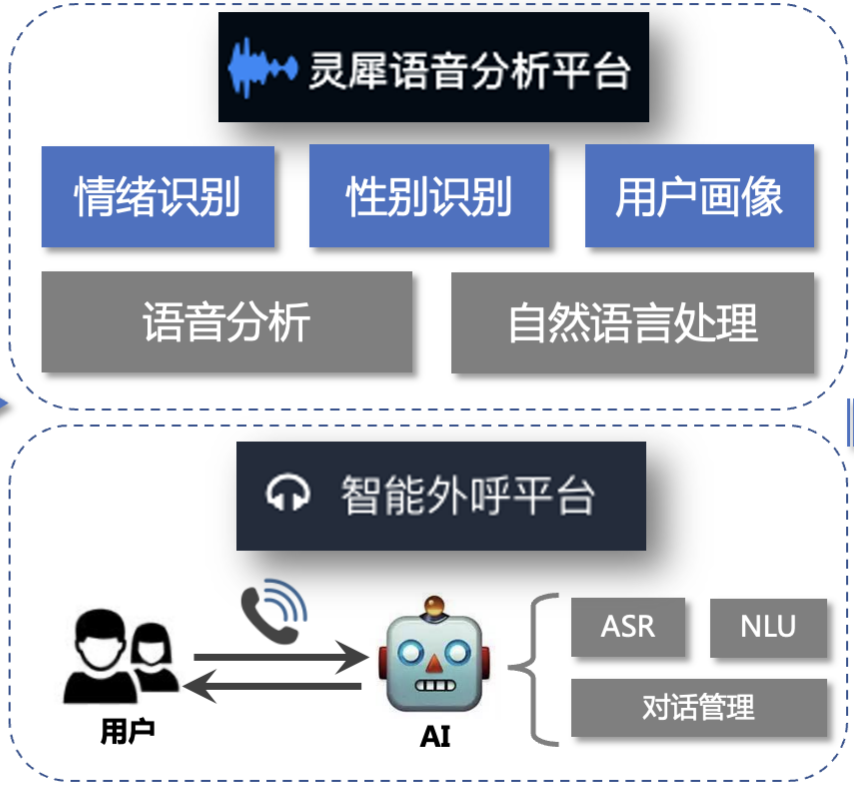
让 生 活 简 单 美 好

电话音频 8K

网络音频 16K



语音识别



业务

智能语音在58的应用背景

让 生 活 简 单 美 好

业务应用

- 神奇面试间
- 语音通知提醒
- 销售客服质检
- AI讲房
- 黄页销售外呼助手
- 语音运维提醒
- 房源信息验真
- 音视频审核

平台产品

- 语音交互平台
  - 多轮对话
  - 电话外呼
  - 来电接起
  - 话术配置
  - 意图识别
  - 对话管理
- 语音分析平台
  - 性别识别
  - 身份识别
  - 意向识别
  - 信息验真
  - 身份虚假识别
  - 语义标签

基础能力

- 语音算法技术
  - 语音识别
  - VAD
  - 说话人分离
  - 语音合成
  - 声纹识别
  - 标点识别
- 语音工程技术
  - 录音文件识别
  - 流式识别
  - 标注系统
- 平台工程技术
  - WPAI
  - WFS
  - WOS

自研语音识别背景

58同城语音识别任务复杂

1. 复杂的业务模式、口音

2. 复杂的通话环境

自研语音识别的价值

更好的识别效果

快速定制能力

BadCase修复
新场景支持
热词增强

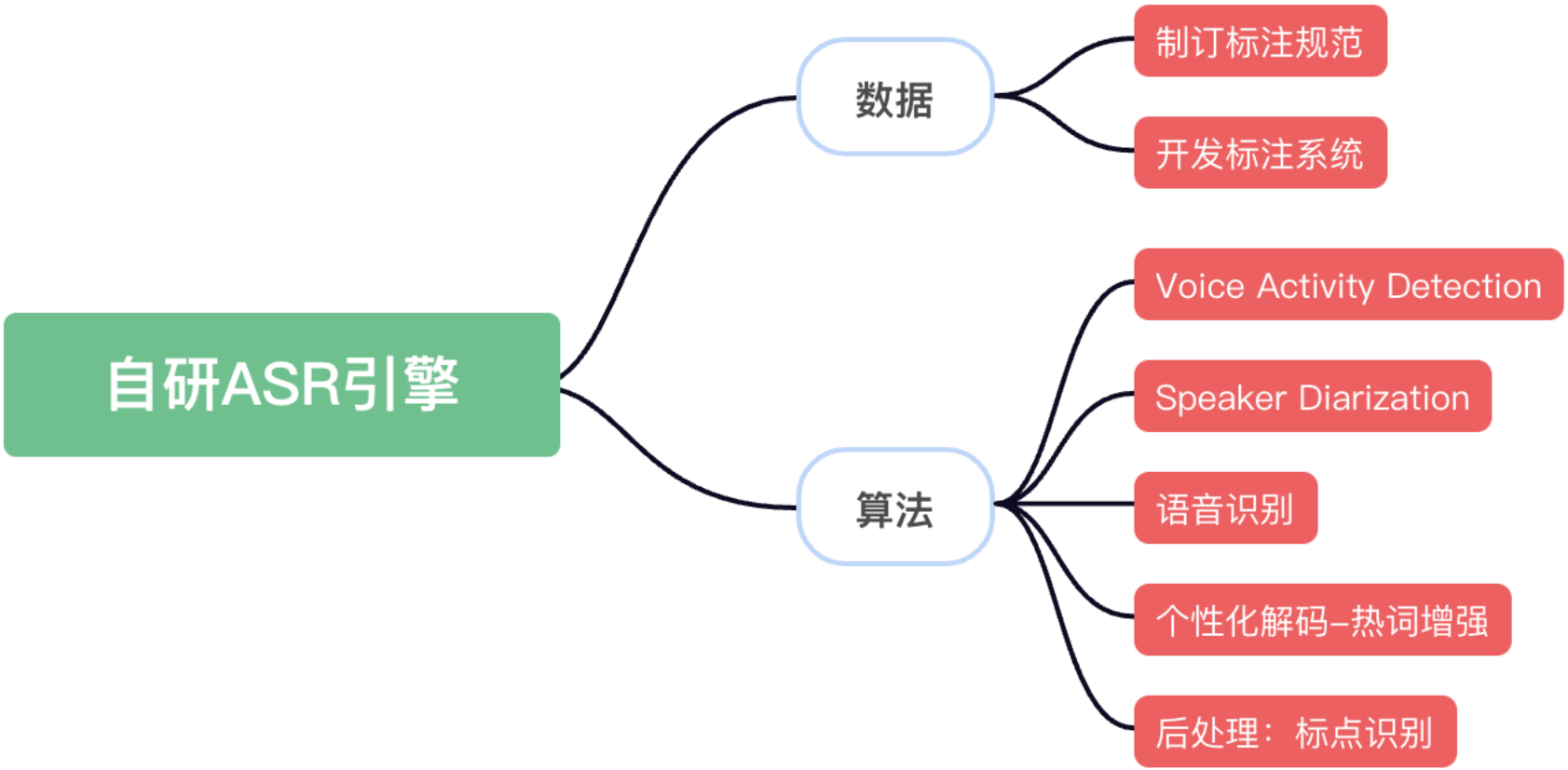
节省成本

比采购第三方节省 数百万元/年

数据隐私

自研语音识别引擎主要工作

让 生 活 简 单 美 好



语音数据标注

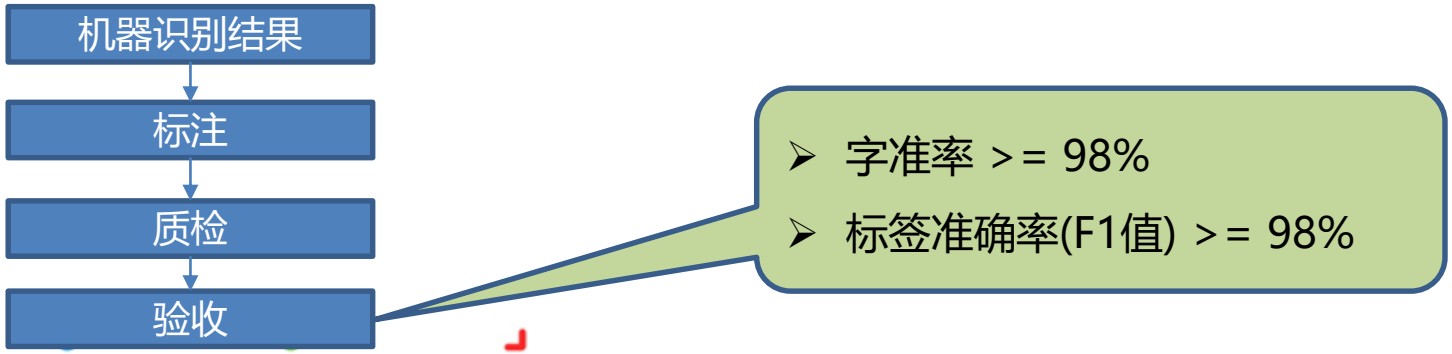
数据特点

- 业务场景繁多：近30种不同业务场景
- 数据来源复杂：电话信道(8k)网络平台(16k)

标注数据选取原则

- 重要业务场景多标注
- 线上效果差的场景多标注

标注流程



语音标注系统

让 生 活 简 单 美 好

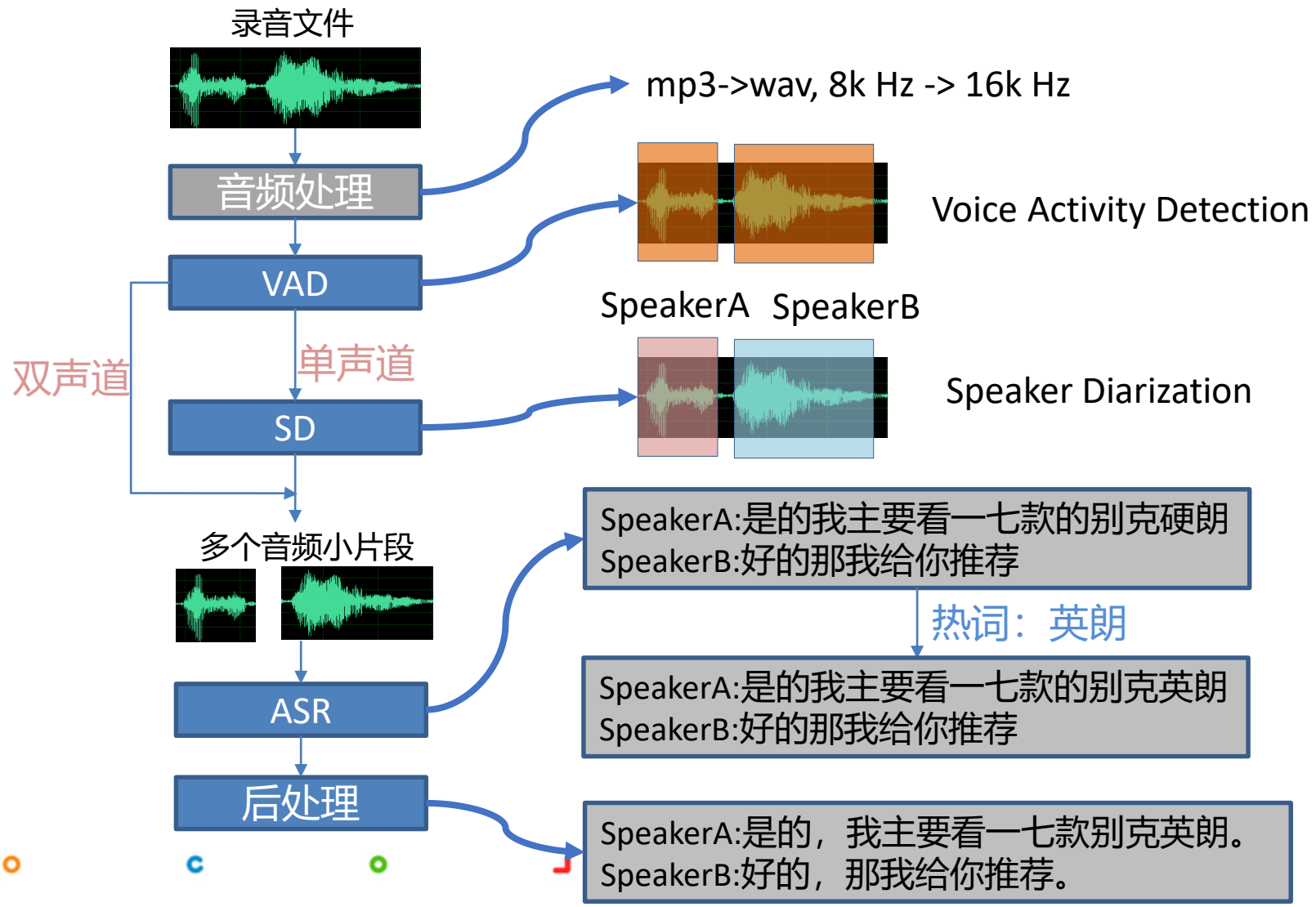
语音标注系统

- Tornado + wavsufer.js实现
- 语音标注、质检、权限管理、数据统计

标注人效：1小时有效音频/每人每天

语音识别引擎算法模块介绍

让 生 活 简 单 美 好



VAD应用场景及功能

流式场景中的VAD

音频信号数据流 → VAD系统 → 当前对话状态：人声开始、持续、结束 → 人机对话系统：根据当前对话状态决定下一步动作

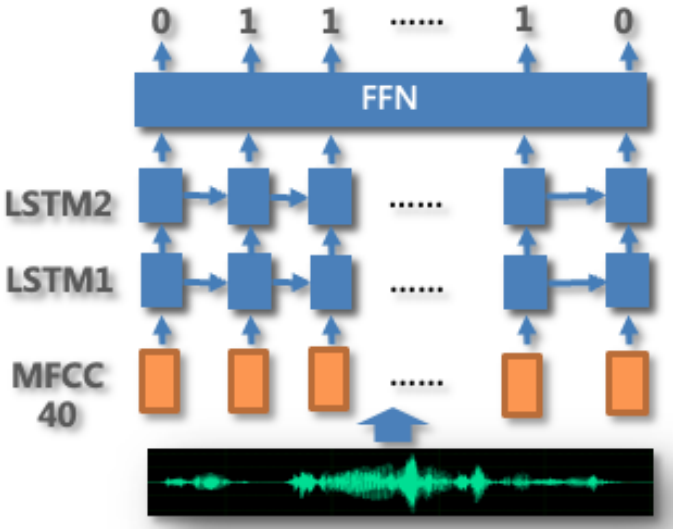
离线场景中的VAD

录音文件
1.mp3
2.mp3
3.mp3
……

→ 离线VAD → Speaker Diarization ASR

VAD模块结构

基于神经网络的决策模块



1:speech 0: non speech

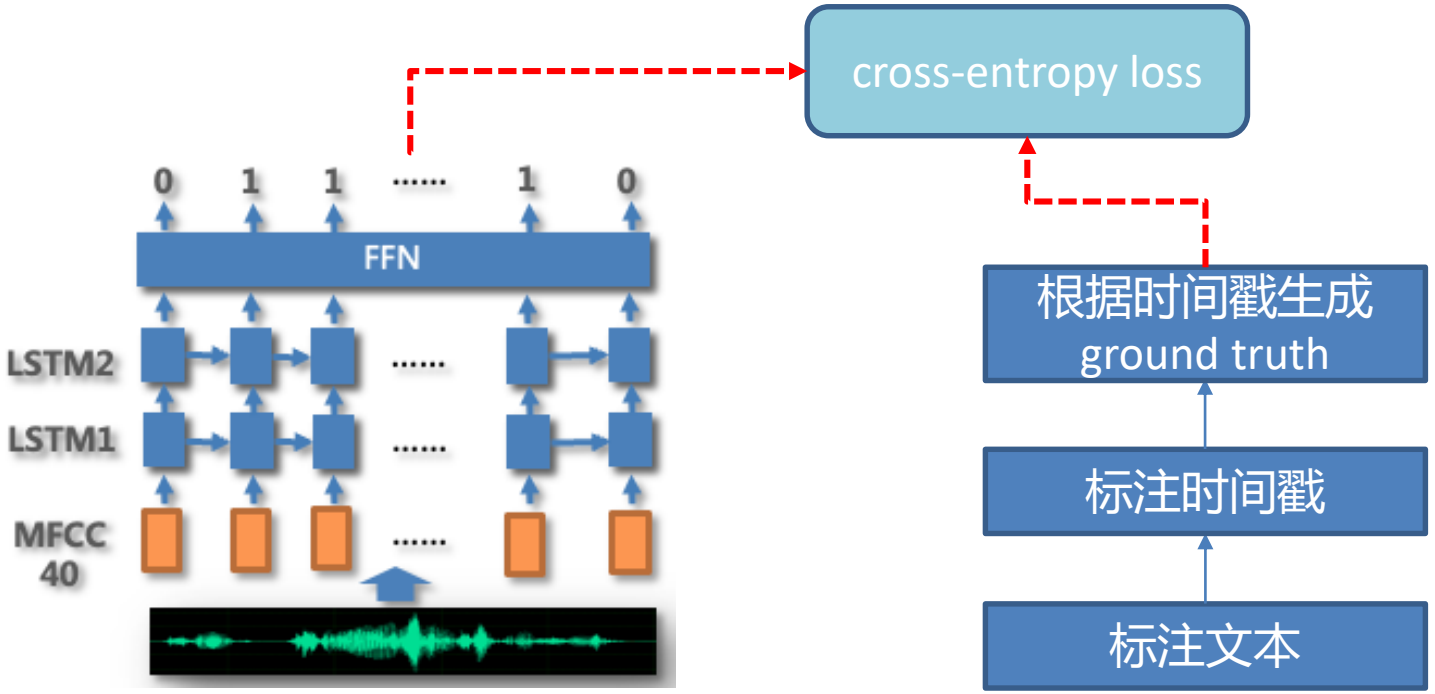
[0,0,0,1,0,0,0,0,0,1,1,0,1,1,1,1,0,1,0,0,0]

基于滑动窗的平滑模块

参数：{N,T1,T2}
规则：N个label的滑动窗中，超过T1个label是1，表示句子开始，超过T2是0，表示句子结束

[0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,0,0,0]

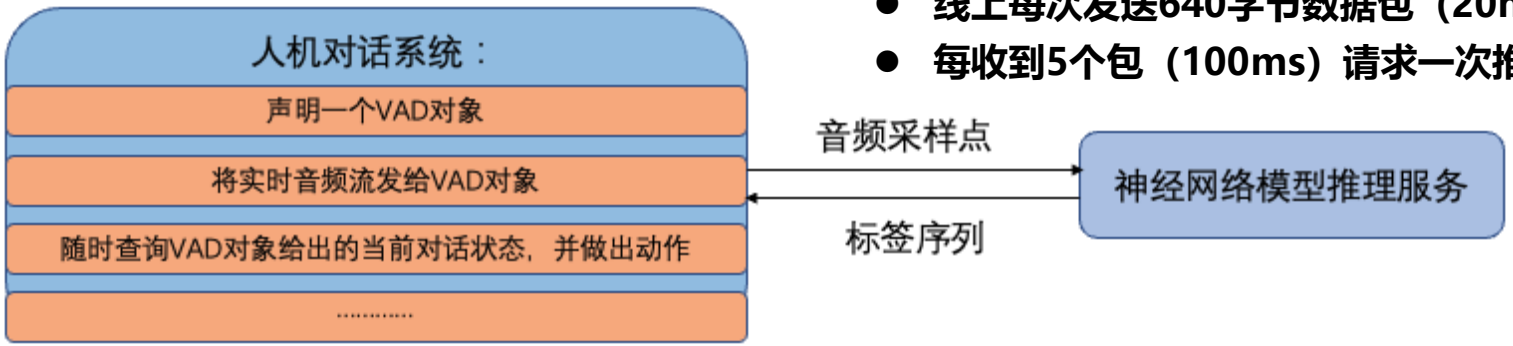
VAD模型训练



- 流式场景使用的模型在训练时LSTM的输入为整条音频
- 离线场景使用的模型在训练时LSTM的输入为变长音频
- VAD模型同样存在domain mismatch的情况，例如用网络音频数据训练的模型在电话场景中效果就很差

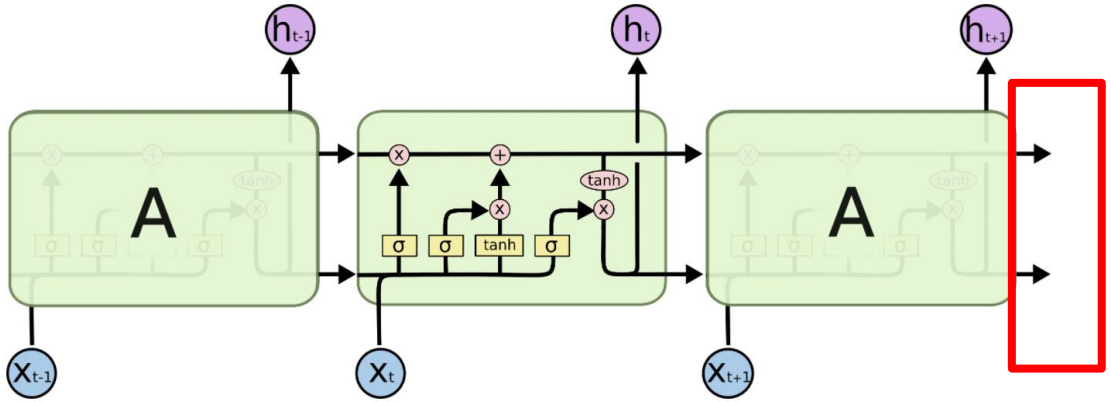
流式VAD的工程化

流式场景调用关系



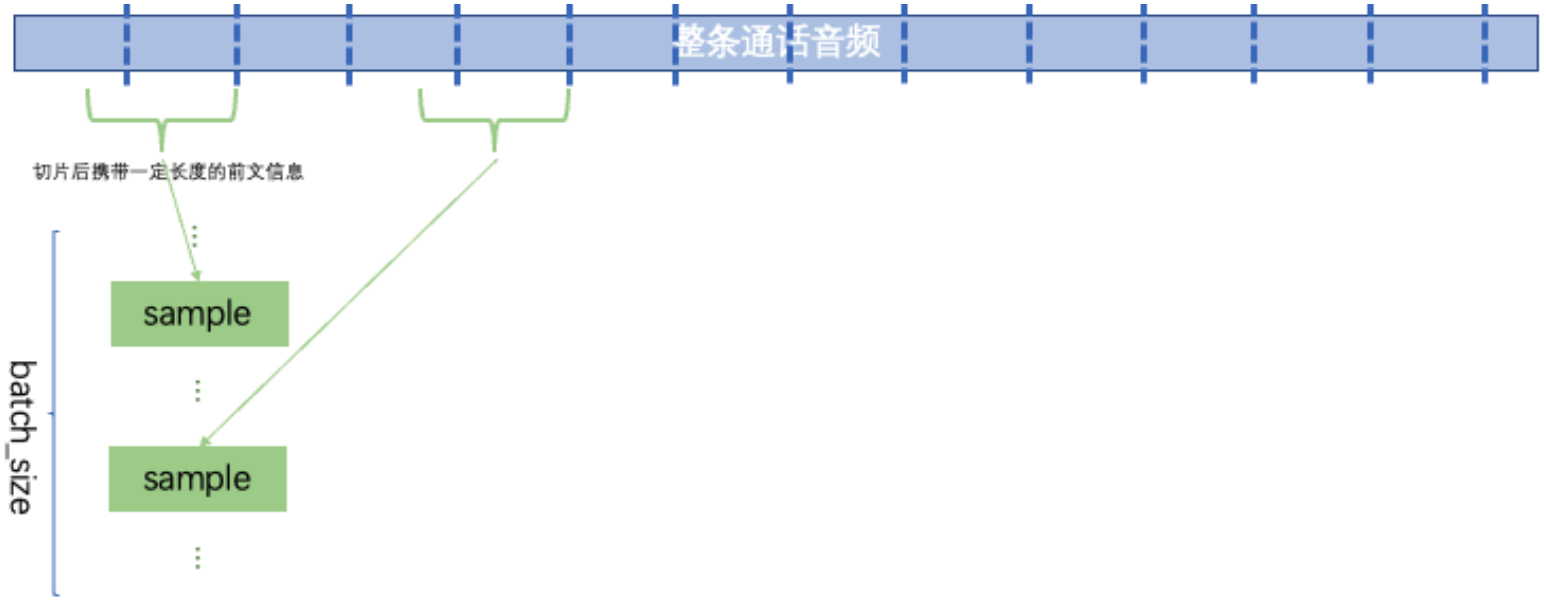
- 线上每次发送640字节数据包（20ms）
- 每收到5个包（100ms）请求一次推理服务

模型历史信息记忆



离线VAD的工程化

切片并行处理

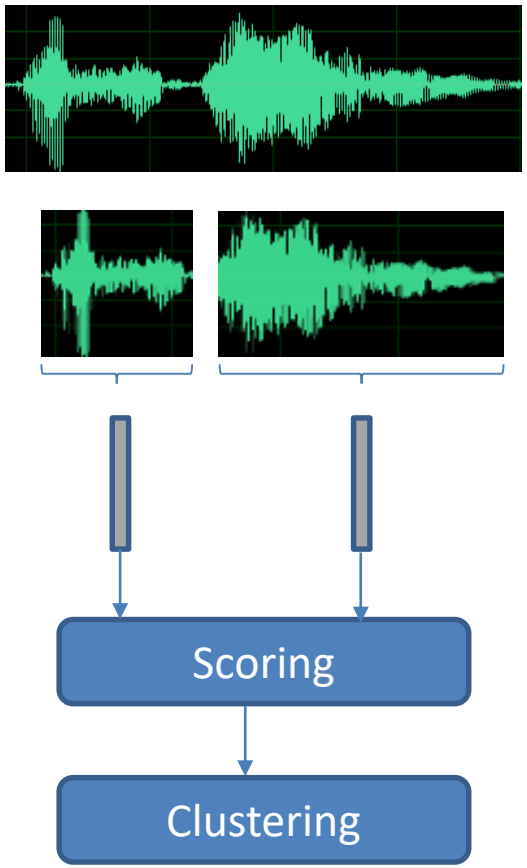


模型短时音频适应

训练样本随机切为3-10秒的片段

Speaker Diarization训练

SD方案：Clustering-based SD



VAD

这里有实验引入说话人变更点检测，der下降5%，但是会增加耗时，所以去掉了这个模块

Speaker embedding extraction

1.考虑到耗时，对VAD结果整句提取embedding，没有用1s窗长0.25s窗移来提取embedding。
2.Embedding extractor用声纹识别训练的resnet，优于ivector

Similarity measurement

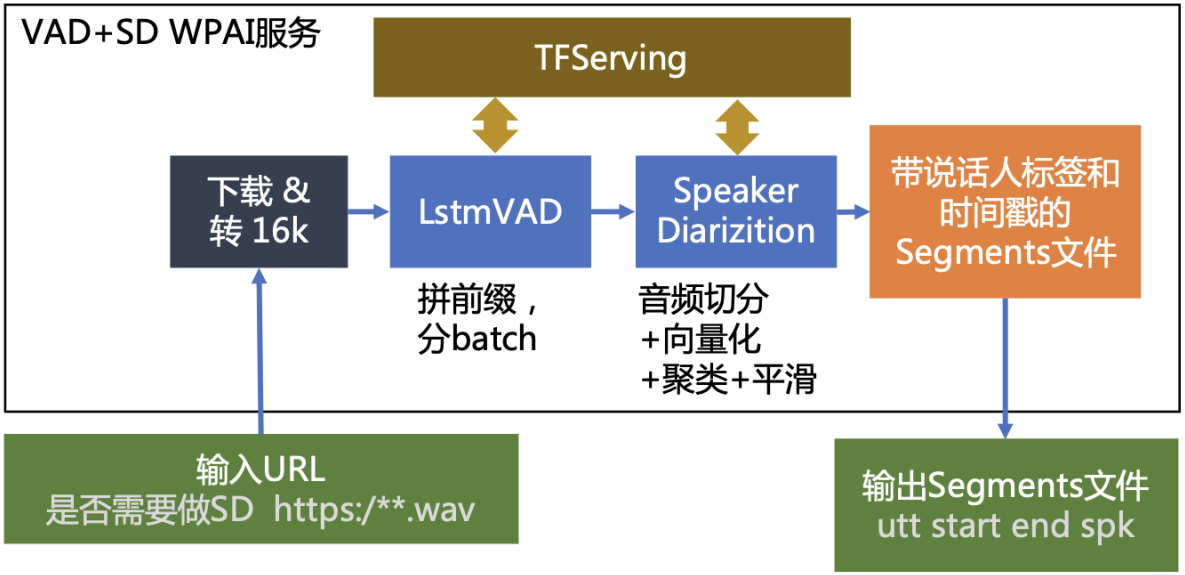
cosine similarity

Clustering

kmeans

Speaker Diarization部署

录音文件识别中VAD和Speaker Diarization部署

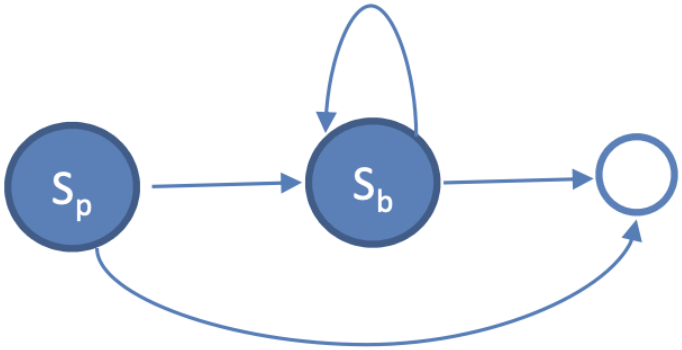


耗时优化

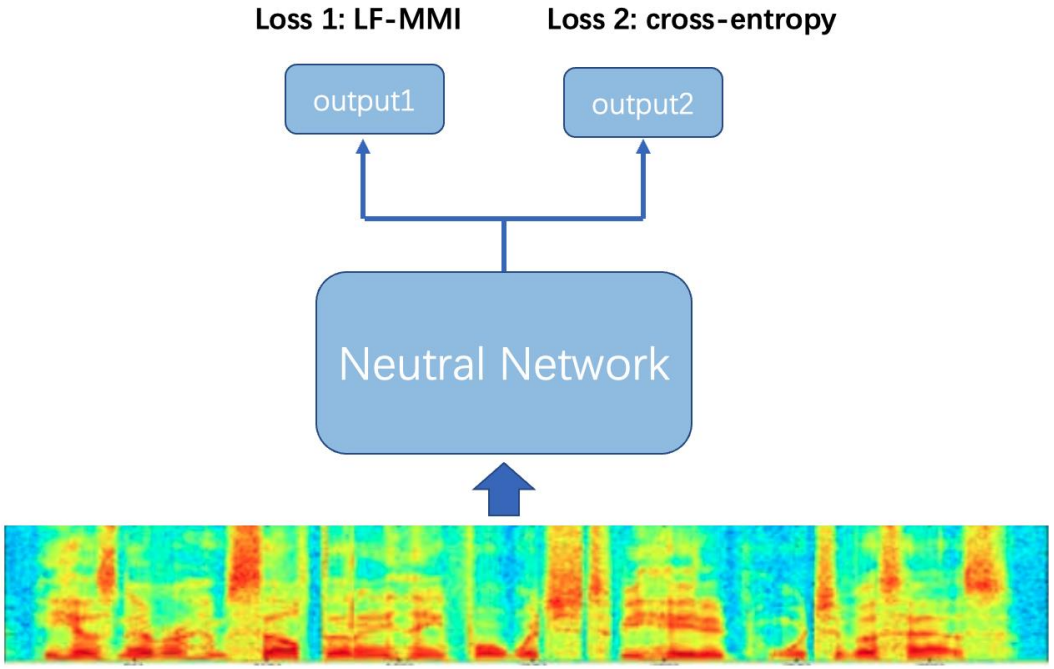
- VAD：24小时 录音/每小时 → 402小时 录音/每小时
- SD：264.54小时 录音/每小时 → 402.57小时 录音/每小时

ASR声学模型

Hybrid ASR System-Chain Model



HMM拓扑结构



神经网络结构

ASR声学模型

Chain Model的训练

- 数据方向
  - 训练数据由所有场景数据混合而成。
  - 单个场景需要至少100h的in-domain data，重点场景300h。
  - 变速、加噪声(场景匹配的噪声)、specaugmentation均有提升。
- 输入特征方向
  - 使用80维fbank，根据训练数据采样率(8k/16k)来灵活调整mel滤波器的上界频率。
  - 全局特征归一化在多场景训练集上表现不好，对单一场景效果好一些，直接用volume perturbation替代归一化操作。
  - 80维fbank的基础上增加100维ivector说话人特征有好处。
- 模型方向
  - cnn+tdnnf的结构好于tdnn。
  - 增加模型的上下文宽度会提升性能。

ASR声学模型

	训练数据量	声学特征维度	Ivector-dim	模型版本	数据增强方式	实验测试集CER
1	200小时	40	0	TDNN	音量	12.53%
2	200小时	40	100	CNN-TDNNF	变速、音量	12.13%
3	200小时	80	100	CNN-TDNNF	变速、音量	11.99%
4	200小时	80	0	CNN-TDNNF	变速、音量、谱增强	11.74%
5	200小时	80	100	CNN-TDNNF	变速、音量、谱增强	11.60%
6	348小时	80	100	CNN-TDNNF	变速、音量、谱增强	11.25%
7	348小时	80	100	CNN-TDNNF	变速、音量、谱增强、加噪声	11.12%

ASR声学模型

Chain Model的训练

➢ 新训练方式的尝试：E2E Chain Model

- 训练流程简单，直接GPU。
- 训练数据需要统一变速到固定的几个长度。
- 建模单元以及决策树最好使用常规Chain Model生成好的。
- 模型效果(CER)差于常规Chain Model 1%。

序号	建模单元	决策树	数据增强	实验测试集CER
Baseline Model: 常规Chain Model				11.12%
1	monophone	无	变速	12.58%
2	biphone	穷举	变速	12.86%
3	biphone	使用常规模型决策树	变速	12.03%
4	biphone	使用常规模型决策树	变速、音量扰动	11.84%

ASR声学模型

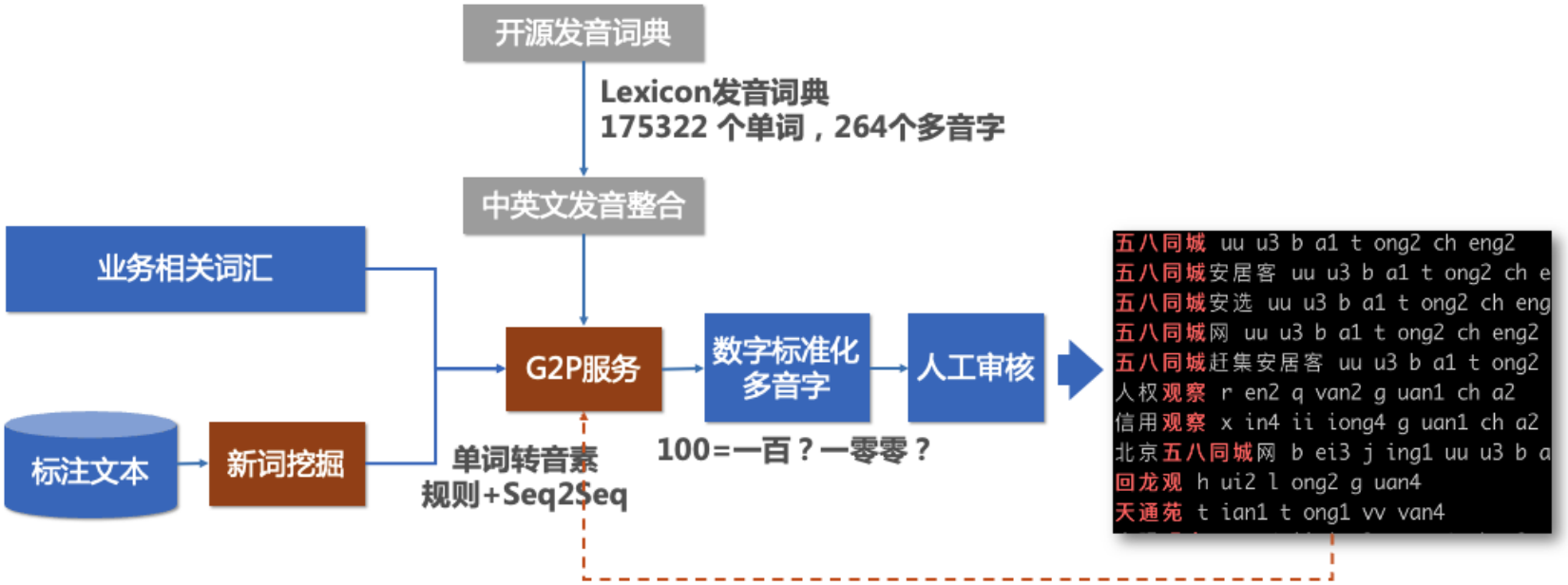
Chain Model的训练

- 新训练方式的尝试：迁移学习
  - 需要用target domain data重新训练hmm-gmm，直接用原来的hmm-gmm对齐效果比较差。
  - 在source domain和target domain差异不大时，NN模型不需要在seed model上添加新的层。
  - Source domain -> target domain迁移学习效果还是不如两组数据加一起从头训练

序号	HMM-GMM方案	神经网络fine-tuning方案	测试集CER
baseline	source domain data混合target domain data直接训练		15.20%
1	不训练HMM-GMM	seed model直接finetuning	17.42%
2	重新训练HMM-GMM	seed model直接finetuning	15.85%
3	重新训练HMM-GMM	seed model新加入输出层	19.56%

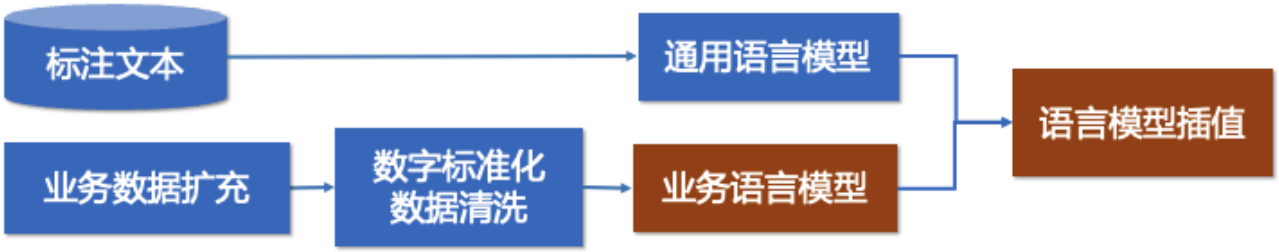
ASR发音词典

自有发音词典构建



ASR语言模型

语言模型优化流程



分词+分字语言模型融合 解决未登录词问题

RNNLM Rescore 带来1%提升

```
40006_415df6f46e712950_right_spkA_m_slice2-1
0 1 <eps>
1 2 家里      18.6107,    -326.687,    2_3216_2422_2
2 3 把        14.3991,    -188.362,    2374_2373_688
3 4 我家      18.4758,    -343.6,      74_73_73_6056
4 5 房子      16.6388,    -250.502,    1886_1885_188
5 6 挂        13.297,     -150.377,    5658_5657
6 7 吧        1.55566,    -87.0326,    5657_5657_704
7
              graph_cost  acoustic_cost
```

ASR语言模型

语言模型语料业务数据扩充

实验序号	声学模型训练数据	语言模型训练数据	电话租房测试集字错误率	二手房测试集字错误率
1	电话租房	电话租房	24.60%	38.00%
2	电话租房	电话租房 微聊房产文本	23.82%	33.47%

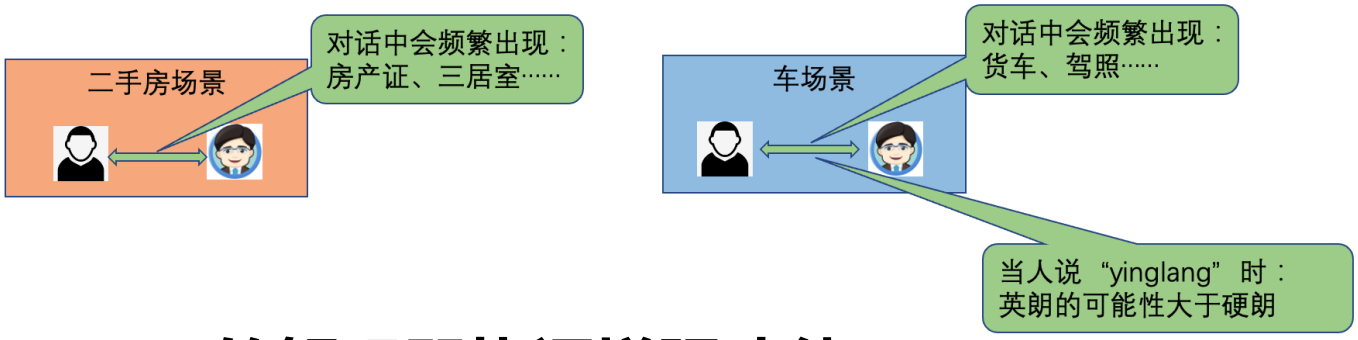
RNNLM Rescore

	场景A测试集	场景B测试集
4-gram一遍解码	13.48%	15.36%
Nbest重打分	12.78%	14.53%
Lattice重打分	12.56%	14.22%

ASR解码器热词功能

热词功能背景

- 通用语言模型对细分场景具有业务特点的词识别率不高



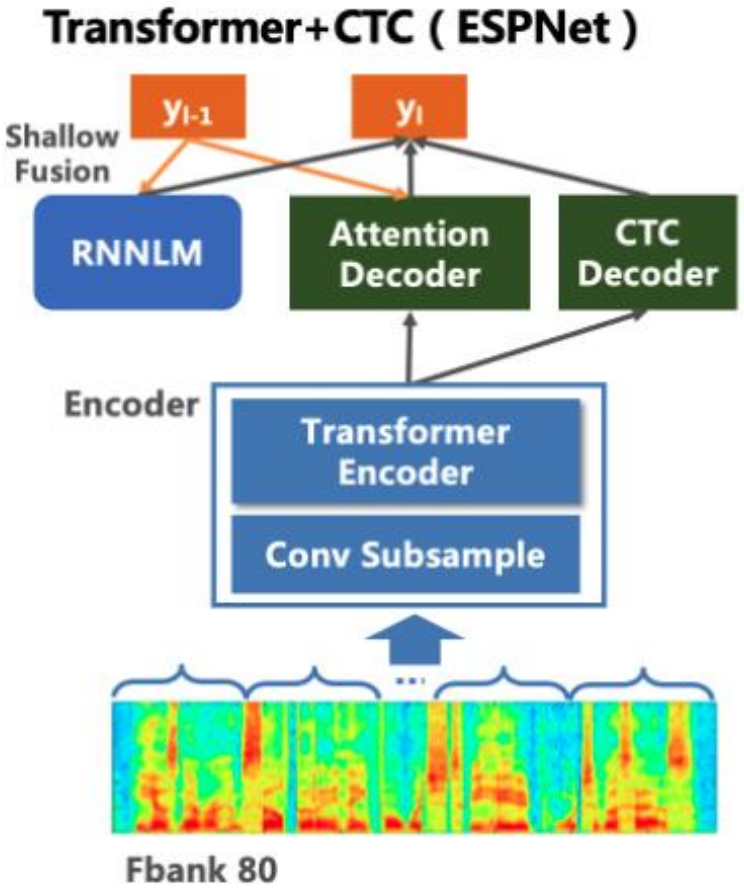
基于WFST的解码器热词增强功能

- 当某条解码路径的弧，其输出词属于热词时，修改这条路径的得分

效果展示

No.	标注文本	热词增强前识别文本	热词	热词增强后识别文本
1	我 没有 招聘 的 需求 我们 现在 人员 饱和	我 没有 消息 需求 我们 现在 人员 饱和	招聘	我 没有 招聘 需求 我们 现在 人员 饱和
2	一七款 英朗	一七款 应 了	英朗	一七款 英朗
3	回访 我 车辆 的 还是 干什么	回访 我 撤掉 了 还是 干什么	车辆	回访 我 车辆 的 还是 干嘛

端到端ASR的探索



效果迭代

- 数据增强：变速+加噪均有效，字准率1-2%提升
- ckpt-average后，效果更好，字准率1-3%提升

解码效率优化

- 批量推理，动态batch_size优化
- 跳帧，更大stride或者更多层的卷积
- 减小beamsize

在测试集上比Chain Model方案的字准率整体高2.25%

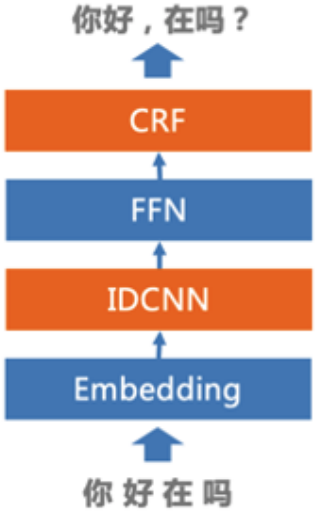
后处理-加标点

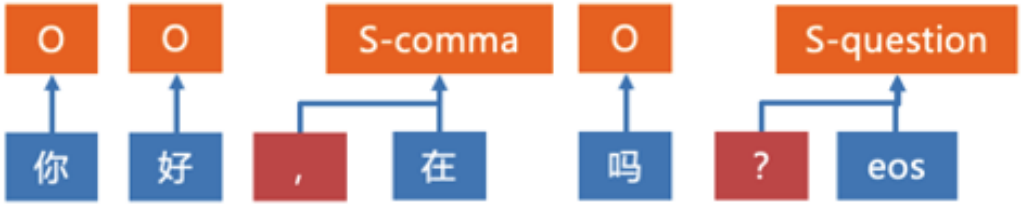
标点识别的作用

- 增加人工质检可读性
- 方便使用规则

加标点前	加标点后
哦现在不要嘛我有个专门的人事在搞这个事情	哦，现在不要嘛，我有个专门的人事在搞这个事情。
好吧唉那咱们年前这段时间会不会加一些人手什么的	好吧，唉那咱们年前这段时间会不会加一些人手什么的？
噢对对对我说这个就是说这边正正好有这个活动嘛哥	噢，对对对，我说这个就是说这边正正好有这个活动嘛哥。

标点识别作为NER任务





ASR效果

让 生 活 简 单 美 好

数据	自研ASR字准率	合作第三方	互联网大厂A	互联网大厂B
所有离线场景汇总	base	-3.01%	-11.46%	-14.82%

- 呼叫中心(电话录音)：招聘客服场景，用户侧字准率81.70%，客服侧字准率91.12%
- 隐私通话(电话录音)：二手房隐私通话场景字准率91.09%

数据	自研ASR字准率	互联网大厂A	互联网大厂B
流式场景A	base	-2.55%	-1.35%
流式场景B	base	-4.52%	-4.3%
流式场景C	base	-2.88%	-0.02%

- 神奇面试间(网络音视频)，字准率91.12%

总结

数据标注

- 持续观察线上效果
- 标注资源向重点场景倾斜

算法

- 拥有语音全链路处理能力：VAD->SD->ASR->加标点
- 拥有个性化定制能力：
  - 重要场景独立优化声学语言模型
  - 解码器热词定制

工程架构部分

王 焱

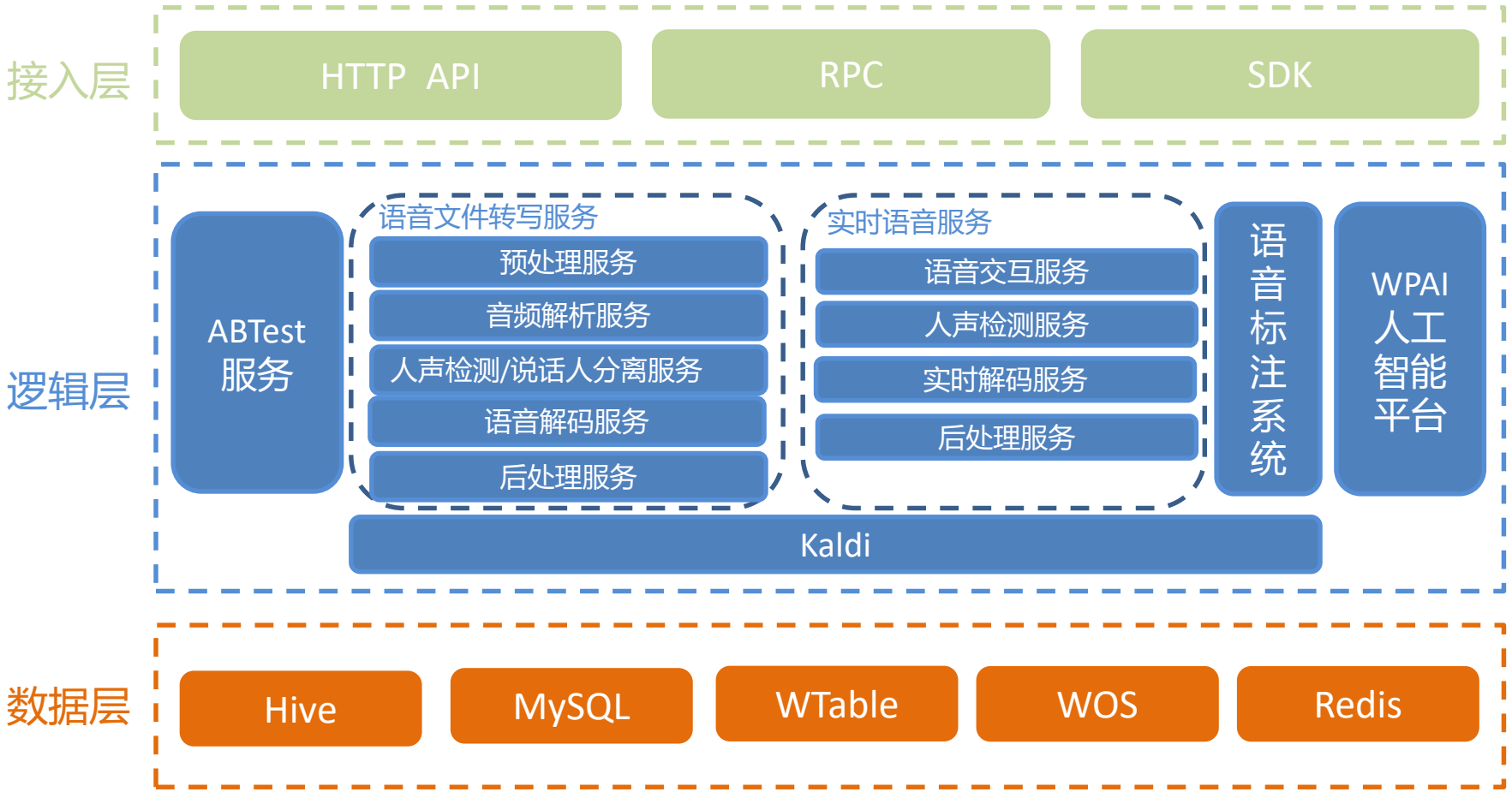
个人简介

- 王焱，58同城AI Lab后端架构师，2017年2月加入58同城，目前主要负责语音识别引擎后端架构设计和开发工作，曾先后负责过推荐系统、智能语音机器人系统的后端架构与开发工作。2012年硕士毕业于华北计算机系统工程研究所，曾就职于Thomson Reuters、H3C等。

- 联系方式：wangyan45@58.com

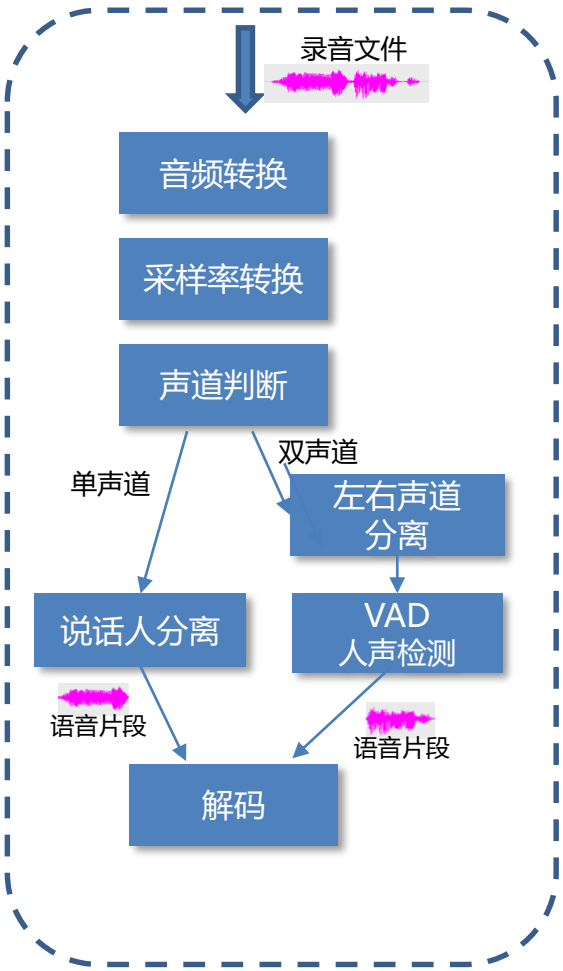
目录

语音识别引擎整体架构

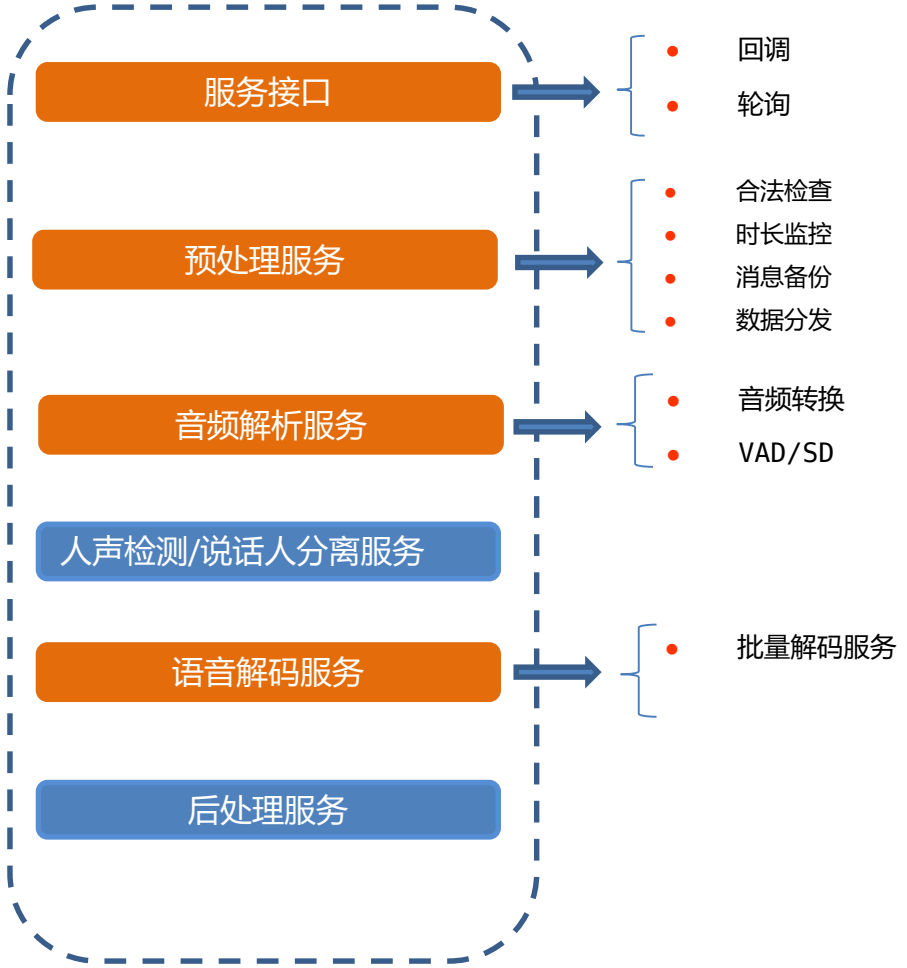

接入层
HTTP API
RPC
SDK
逻辑层
ABTest 服务
语音文件转写服务
预处理服务
音频解析服务
人声检测/说话人分离服务
语音解码服务
后处理服务
实时语音服务
语音交互服务
人声检测服务
实时解码服务
后处理服务
语音标注系统
WPAI 人工智能平台
Kaldi
数据层
Hive
MySQL
WTable
WOS
Redis

语音文件转写服务

录音转写的一般流程：

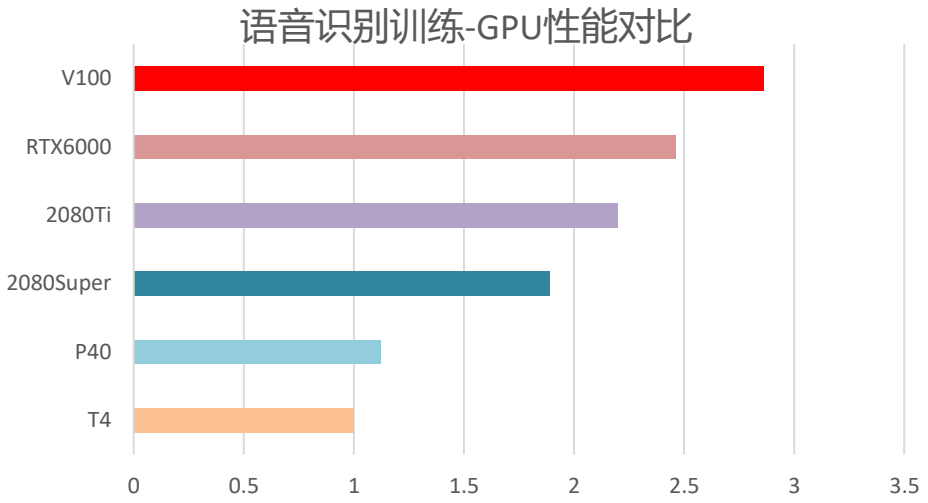


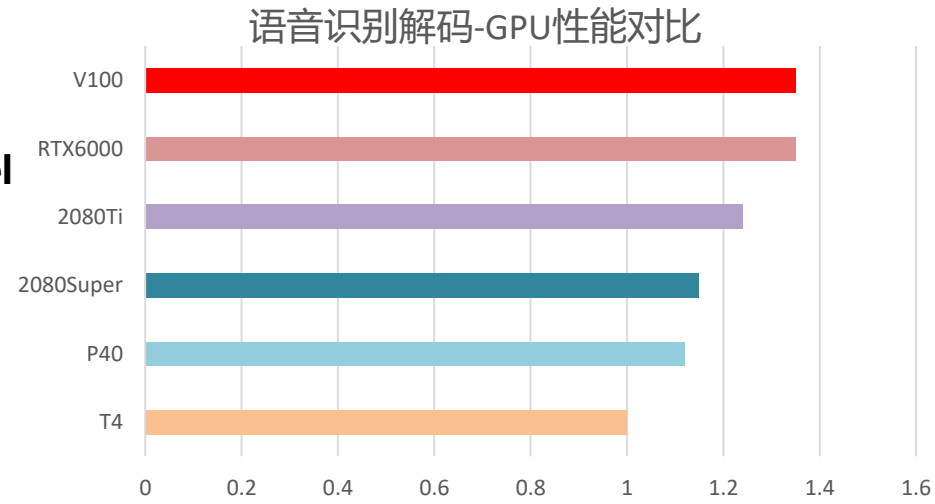
语音文件转写服务：



解码器和解码效率

- Tesla V100
- Quadro RTX 6000
- GeForce RTX 2080Ti
- GeForce RTX 2080super
- Tesla P40
- Tesla T4

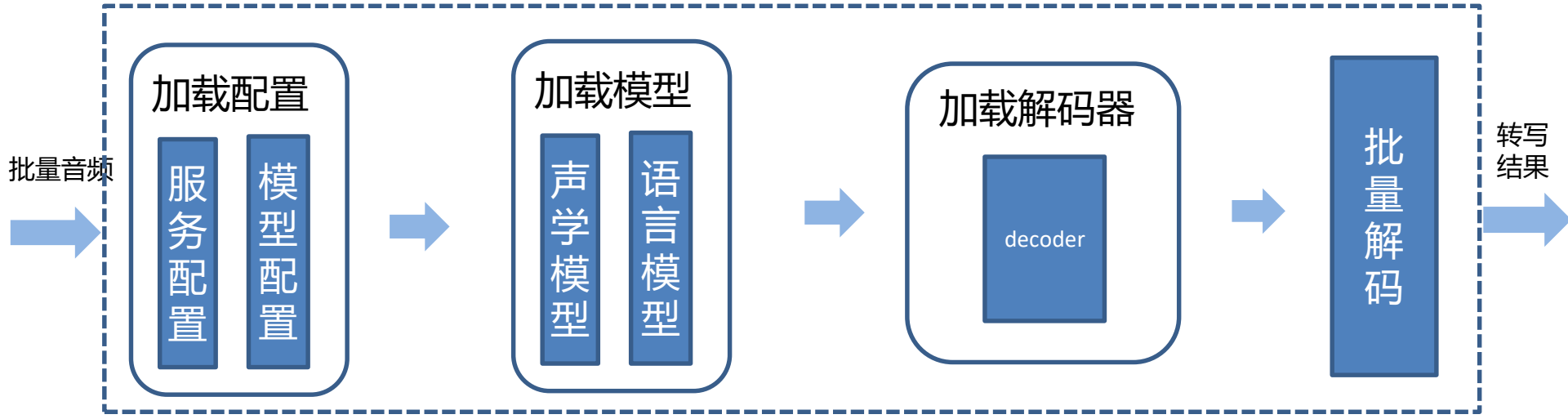




- 选择的解码器
  - CPU解码器：**nnet3-latgen-faster-parallel**
  - GPU解码器：**batched-wav-nnet3-cuda**

批量解码服务的实现

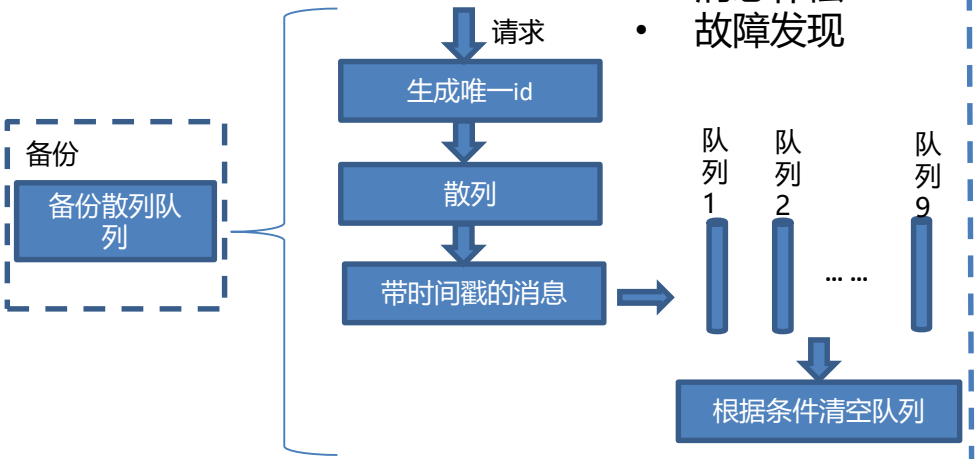
- 解码器服务化：
  - 基于batched-wav-nnet3-cuda/nnet3-latgen-faster-parallel改造
  - 单体应用，可扩展性、可维护性
  - 灵活部署和扩展、便于监控
  - 探索在细粒度的资源部署和运行
  - 主要问题：编译问题/运行时问题/不同显卡运行模式下服务能力
  - 测试结论：和离线对比，CER一致；耗时小于等于离线解码



服务优化实践

可靠性保证

- 消息备份
- 消息补偿
- 故障发现

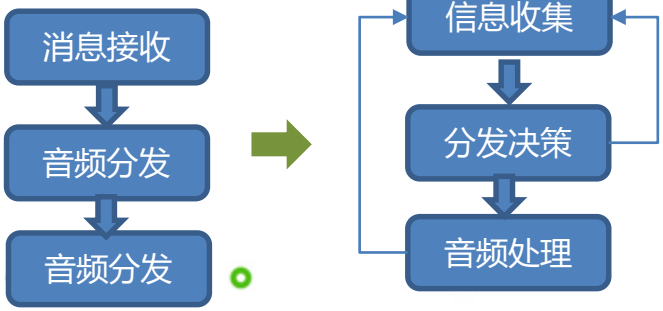


在全链路上降低耗时

- 并行处理
- 减少阻塞、等待时间

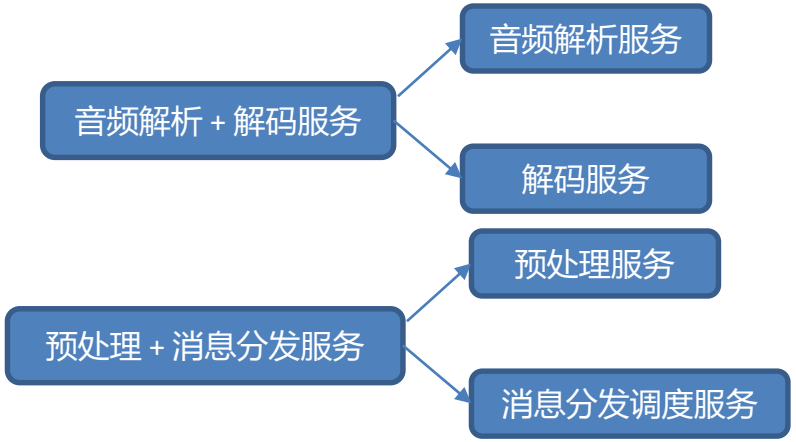
优化资源使用效率

- 资源隔离到资源共享
- 均衡、提高资源利用率
- 降低解码空闲等待时间



服务解耦

- 服务按功能横向拆分，独立迭代，独立优化

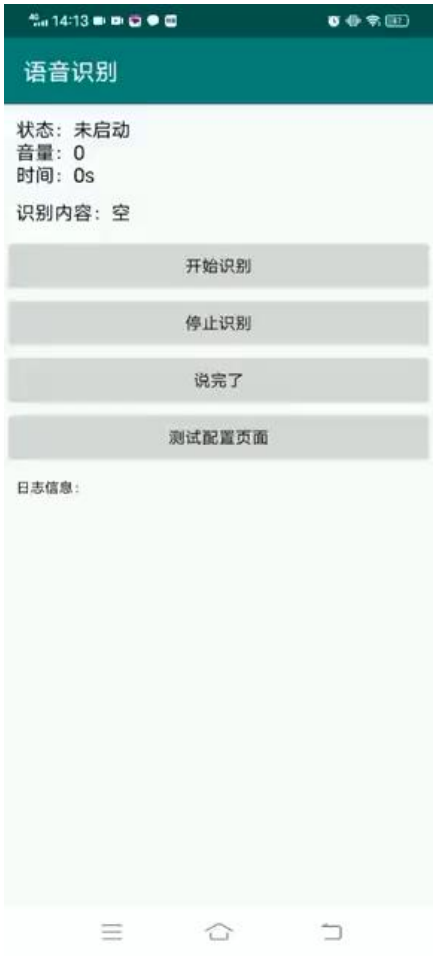


服务的应用

- 已经接入了30个业务场景，包括销售、客服、隐私通话、微聊等，日均处理近2万小时
- 部署GPU 31台机器(62张卡)，CPU 32台机器
- 目前服务部署在docker里，自动化工具做部署，后续迁移到k8s平台统一管理

实时识别服务

让 生 活 简 单 美 好



语音数据流

鉴权

限流

VAD
人声检测

实时语音解码

解码器

中间结果

你好

你好请问

你好请问房子

你好请问房子租出去了吗

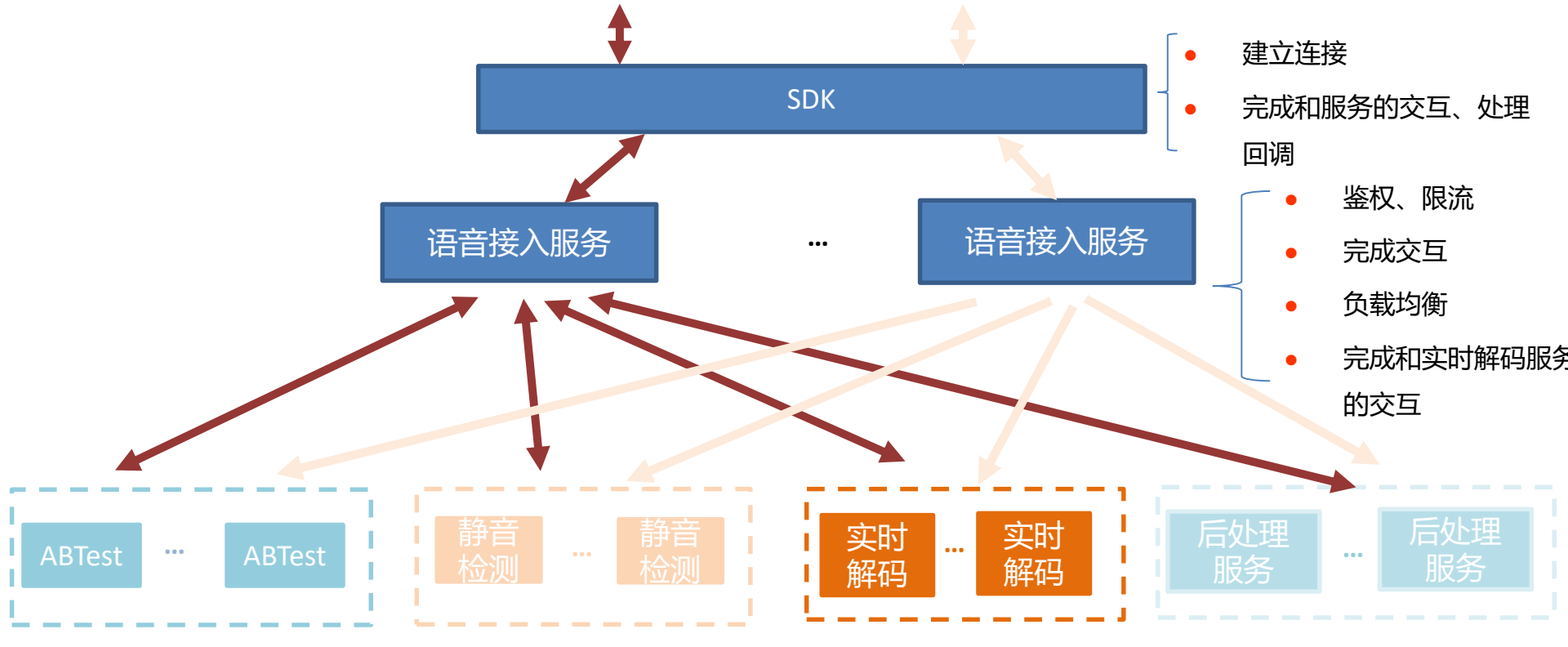
最终结果

你好请问房子租出去了吗

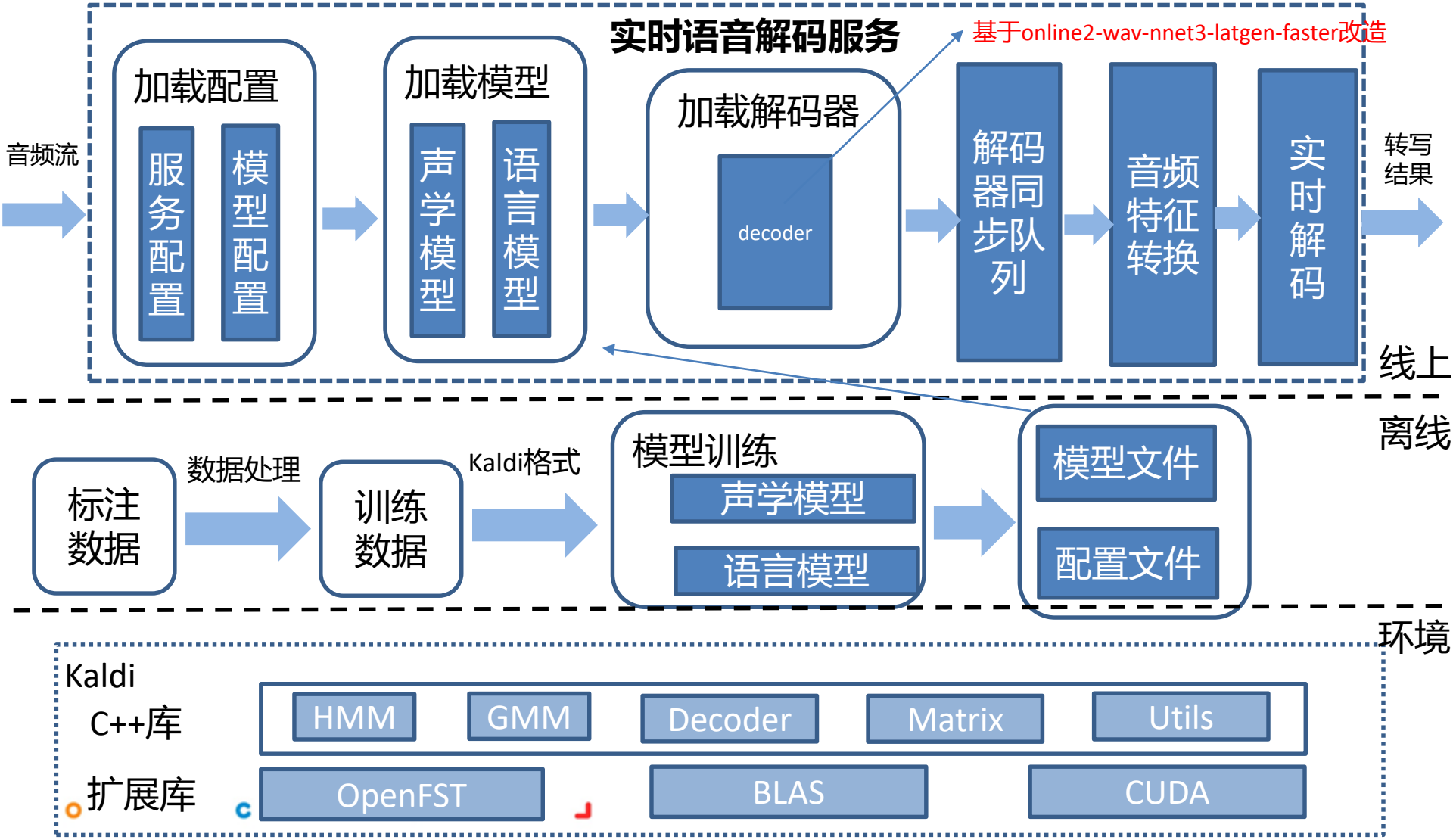
后处理

你好，请问房子租出去了吗?

实时语音识别的实现


SDK
建立连接
完成和服务的交互、处理回调
语音接入服务
...
语音接入服务
鉴权、限流
完成交互
负载均衡
完成和实时解码服务的交互
ABTest
...
ABTest
静音检测
...
静音检测
实时解码
...
实时解码
后处理服务
...
后处理服务

实时解码服务的实现


实时语音解码服务
基于online2-wav-nnet3-latgen-faster改造
音频流
加载配置
服务配置
模型配置
加载模型
声学模型
语言模型
加载解码器
decoder
解码器同步队列
音频特征转换
实时解码
转写结果
线上
离线
标注数据
数据处理
训练数据
Kaldi格式
模型训练
声学模型
语言模型
模型文件
配置文件
环境
Kaldi
C++库
HMM
GMM
Decoder
Matrix
Utils
扩展库
OpenFST
BLAS
CUDA

解码器服务化

原有问题：

- 原生解码器(online2-wav-nnet3-latgen-faster)并发能力很有限
- 单体应用，可扩展性差，可维护性差

优化方案：

- 服务化，使用多线程、多解码器，支持并发请求

效果：

- 提供并发解码能力，多节点部署和扩展的服务

并发的优化

原有问题：

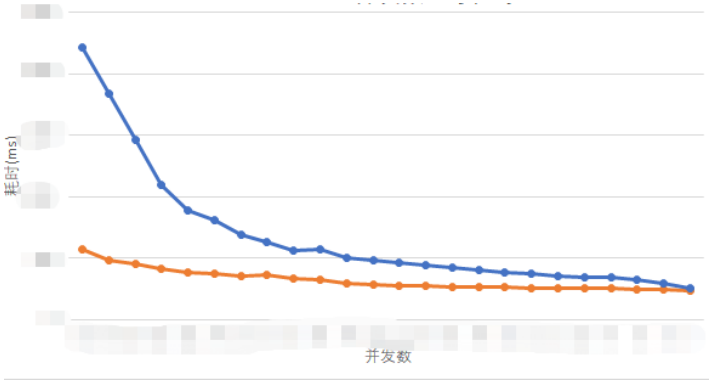
- 耗时随并发升高，最高能接受16路并发

优化方案：

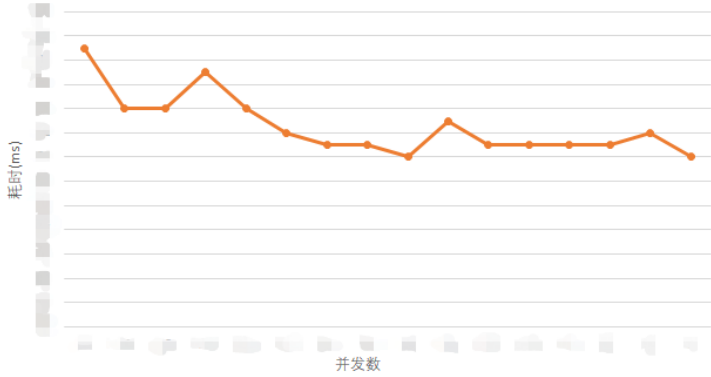
- 优化beam参数
- 优化内存

效果：

- 解码并发数和耗时没有关联
- 从最高16路并发，增加到32路并发
- 相同测试集下对比优化前后流式解码耗时，由567ms降低至97ms



并发数与耗时的关系



并发数与耗时的关系

网络收发包问题的优化

原有问题：

- 客户端sdk丢包、重复包问题
- 服务端收不到、延迟收到调用方发送的数据帧

解决方案：

- 发送队列引用的内存未发送的被覆盖(丢包、重复包)
- nginx收发包问题，临时使用VIP直连后端集群方案

效果：

- 问题解决，无异常，连接稳定

耗时的优化

原有问题：

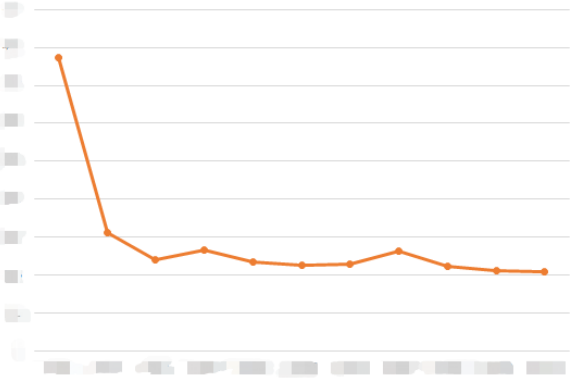
- 解码服务，开启获取中间结果开关后，导致下一帧解码耗时累积增加
- 语音交互服务GC导致耗时高、客户端SDK耗时高

解决方案：

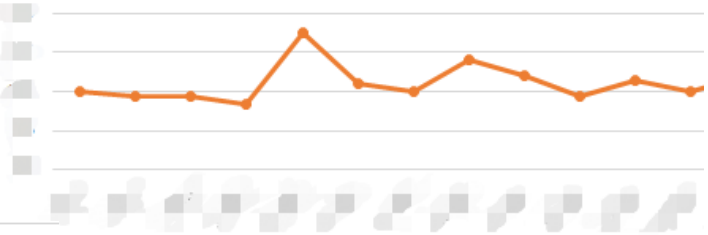
- 优化解码服务，从Lattice选取路径的逻辑
- 优化语音交互服务，优化代码逻辑，优化JVM参数

效果：

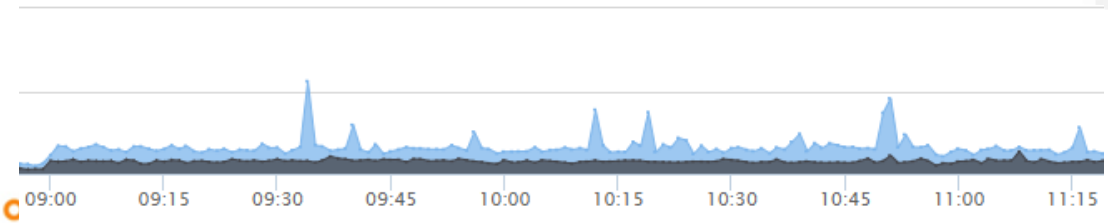
- 打开获取中间结果的开关，不影响下一帧的解码耗时
- 获取中间结果平均耗时从优化前104ms降低至18ms
- 语音交互服务，TP99耗时降低300ms+



一次发送时长与耗时的关系



一次发送时长与耗时的关系



GC优化前后耗时对比

总结

- 数据采集
  - 持续观察线上效果
  - 标注资源向重点场景倾斜
- 算法模型
  - 拥有语音全链路处理能力：VAD->SD->ASR->加标点
  - 拥有个性化定制能力：
    - 重要场景独立优化声学语言模型
    - 解码器热词定制
- 工程架构
  - 语音识别的架构
  - 语音文件转写服务的实现与优化
  - 流式识别服务的实现与优化

欢迎关注

让 生 活 简 单 美 好

欢迎关注 58AILab 公众号

开源项目：

- **《开源 | qa_match：一款基于深度学习的问答匹配工具》**
  https://github.com/wuba/qa_match
- **《开源 | dl_inference：通用深度学习推理服务》**
  https://github.com/wuba/dl_inference

相关文章：

- 3人半年打造语音识别引擎——58同城语音识别自研之路
- **流式和离线场景下VAD语音端点检测算法实践**
- **人物|王焱：58同城流式语音识别引擎应用实践**

58AILab小秘书
北京 朝阳

招聘工程师
欢迎投递 **zhankunlin@58.com**
或加小秘书微信号咨询：WubaAILab

Thanks!

让 生 活 更 简 单