实时识别（新版）

部署文档

版本 V1.0.1

更新时间：2020 年 03 月 17 日

商标声明

注意

您购买的产品、服务或特性等应受腾讯公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，腾讯公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

修订记录

修订日期	修订版本	修改描述	作者
2020 年 02 月 13 日	V1.0.0	[制定初稿]	liquansun
2020 年 03 月 17 日	V1.0.1	完善部署流程	vectorli

修订流程

对于本文档中任何内容的增删改以及相关其它文档的创建，都应该知会测试人员以及相关的干系人。

接口人

本文档中的任何信息都应该被仔细的阅读。如果有任何疑虑，意见或问题，请直接联系下表中的接口人。

姓名	邮箱	电话	组织

目录

1．部署准备

1.1 软件要求

系统要求	查询指令
centos 系统>7.2（装上开发工具）	cat /etc/centos-release
cuda8.0/9.0	cat /usr/local/cuda/version.txt
Nvidia 驱动	nvidia-smi 安装好对应驱动
gcc 4.8.5	gcc -v

1.2 硬件要求

配置项	配置要求	查询指令
CPU	支持 avx2 指令 56 个逻辑核及以上	是否支持 avx2: cat /proc/cpuinfo | grep avx2 查看逻辑核： cat /proc/cpuinfo| grep "processor"| wc -l
GPU	Nvidia，P4*2	nvidia-smi
内存	64G 及以上	free -h
硬盘	200G 以上	df -h（分区在要安装的部署包下，保证 200G 以上）

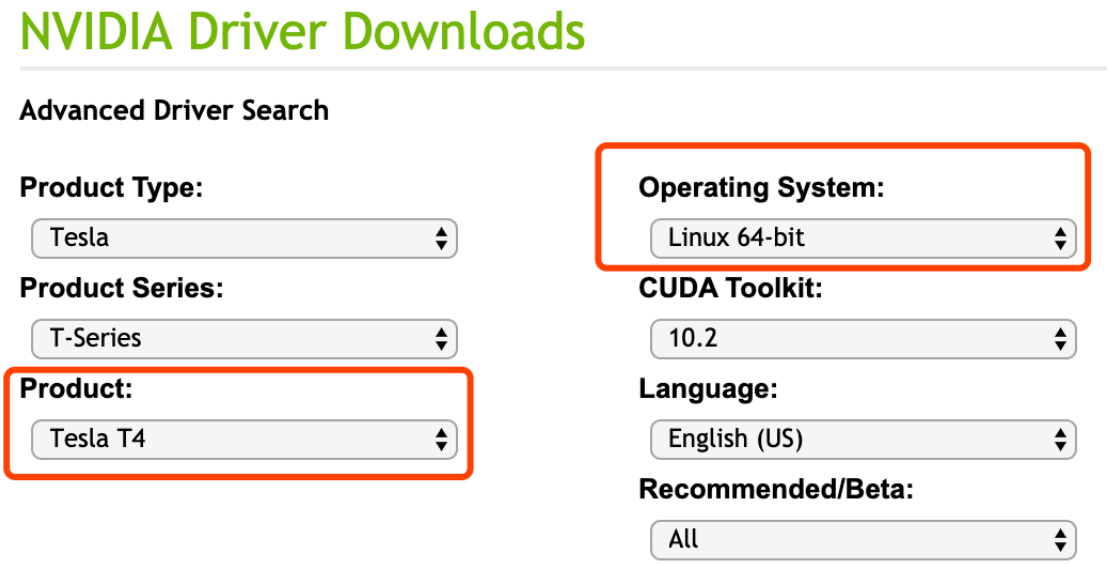
推荐配置：E5-2680v4*2/16G*8/300G*12/10GE*2/P4*2/

需要预先安装好 nvidia 显卡驱动.

可以使用安装包 tools 下的 precheck.sh 检查软/硬件是否满足需求

驱动安装，需要去官网根据显卡和操作系统选择驱动。

https://www.nvidia.com/Download/Find.aspx



安装过程参考：

https://cloud.tencent.com/document/product/560/8048#deb.2Frpm-.E5.8C.85.E5.AE.89.E8.A3.85

安装成功可以使用 nvidia-smi 查看

1.3 系统设置

修改系统最大打开文件数，先查看系统最大打开文件描述符：

ulimit -n

如果数量太低（默认 1024），修改系统最大打开文件数（若大于 65535 则可不用修改）

sudo vim /etc/security/limits.conf

在文件末尾添加：

** hard nofile 65535*

** soft nofile 65535*

1.4 申请授权

在安装部署前，用户机器硬件不变的情况下，运行下载包中 tools/getinfo 下 get_machine_info.sh 获取硬件指纹信息，请发邮件申请授权，获取到 hardwareidlist.bin，qcloudauth.bin ， POC 版本还需要申请加密狗。授权申请，参考：《ASR 授权申请方式说明》

2. 安装包说明

文件名	功能说明
auto-build-asr.tar.gz	部署依赖包。如果机器已经安装 gcc/g++，并且版本在 4.8.5 以上，不需要再安装依赖包。安装依赖包需要 root 权限，只需安装一次
tools.tar.gz	工具包。包含：环境检查、获取硬件指纹等
asr_realtime_*.*.*.tar.gz	服务引擎包。分为加密狗版本和无加密狗版本，安装前需要确认使用的环境
env_***_16k.tar.gz env.online.8k.***.tar.gz	模型包。分为 8k、16k，需要根据使用场景决定下载哪个

说明：服务安装真正需要的只需要服务引擎包和模型包，可以使用普通用户，安装在当前用户下

3. 安装步骤说明

满足实时识别软硬件要求，并行已经安装驱动和 g++，下载服务引擎包和模型包

3.1 下载安装包

以引擎 2.0.0 版本和 8k 模型为例。

下载：asr_realtime_2.0.0.tar.gz、env.online.8k.lstm.lfr.common.v1.4.20190808.tar.gz 放到安装目录

解压：

tar zxvf asr_realtime.tar.gz

tar zxvf voice_rec_engine.tar.gz

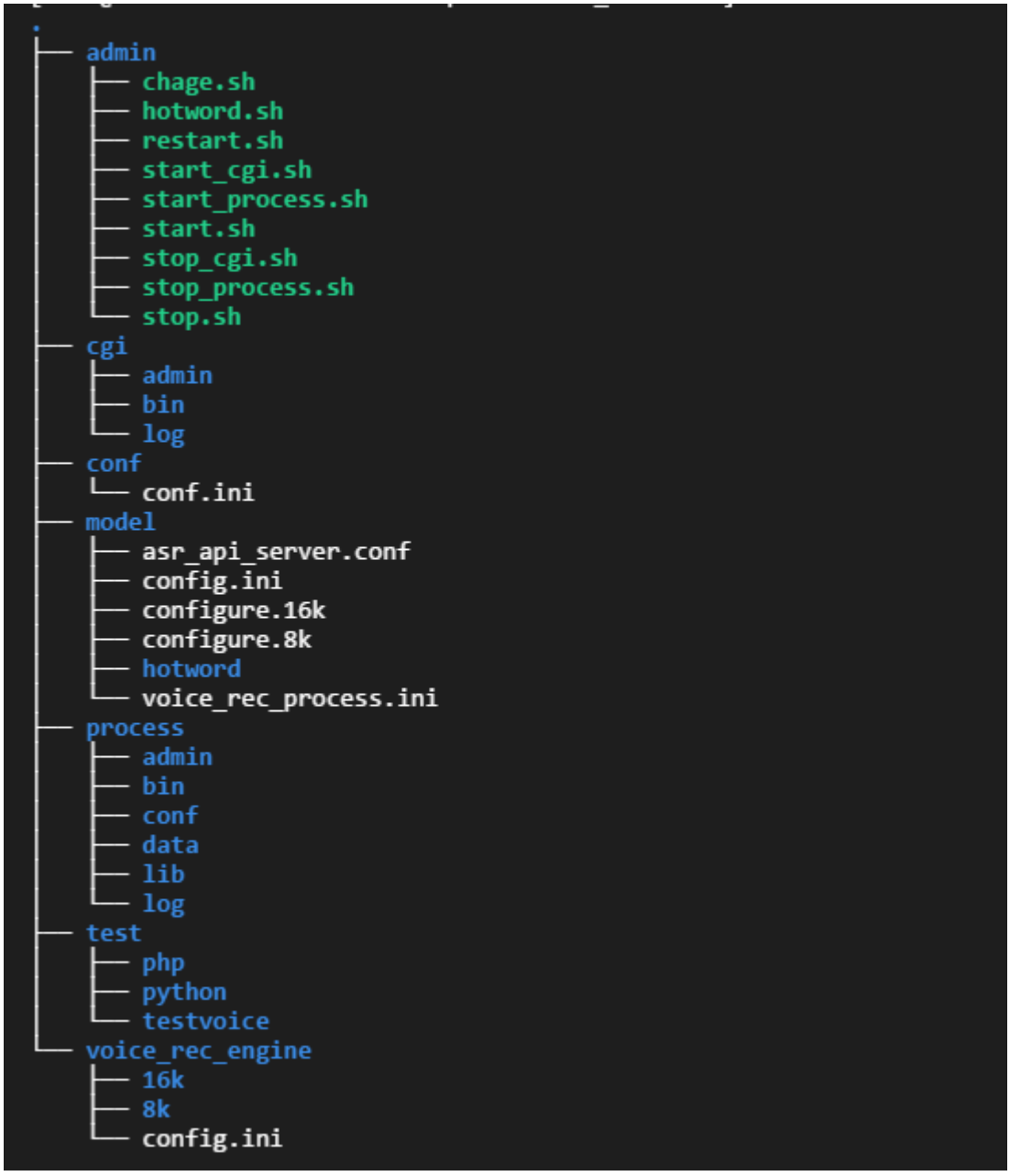
将解压后得到的 voice_rec_engine 文件夹移动入解压后得到的 asr_realtime 文件夹中

mv voice_rec_engine ./asr_realtime

进入 asr_realtime 中

cd asr_realtime

整个目录结构：



目录说明：

文件名	说明
admin	主要存放启动脚本 start.sh 为启动服务脚本 stop.sh 为停止服务脚本 restart.sh 为重启服务脚本 start_cgi.sh 为单独启动 cgi 层脚本 stop_cgi.sh 为单独停止 cgi 层脚本 start_process.sh 为单独启动 process 层脚本 stop_process.sh 为单独停止 process 层脚本 chage.sh 为修改配置使之生效脚本。
cgi	接入层服务

	admin 为接入层服务相关脚本 bin 为服务程序和配置文件 log 为接入层服务日志
conf	实时识别服务配置文件，一般只需要修改此文件即可完成服务正常配置
model	配置文件模板存放地
process	识别层服务存放处 admin 为识别层服务相关脚本 bin 为服务可执行程序存放处 conf 为识别层配置文件存放处 lib 为所需资源链接库 log 为识别层服务日志
test	测试文件夹，testvoice 为可测试音频，提供 php 和 python 两种语言测试 demo
voice_rec_engine	识别模型，分为 16k 模型和 8k 模型。实际部署时只会提供一种

3.2 替换授权文件

将获取的到鉴权文件 hardwareidlist.bin，qcloudauth.bin 放在指定目录。
若部署服务为 16k 识别引擎，将授权文件放在 voice_rec_engine/16k/envs 下
若部署服务为 8k 识别引擎，将授权文件放在 voice_rec_engine/8k/envs 下

```
[root@VM_0_17_centos ~/asr_realtime/voice_rec_engine/16k/envs]# ll
total 20
-rw-r--r-- 1 root root   32 Feb 24 15:23 hardwareidlist.bin
drwxr-xr-x 2 root root 4096 Jan 16 11:27 main
-rw-r--r-- 1 root root  257 Feb 24 11:48 qcloudauth.bin
-rwxr-xr-x 1 root root 3177 Jan 16 11:30 sr.conf
drwxr-xr-x 6 root root 4096 Feb 13 11:15 textprocess_env
```

3.3 修改配置文件

vim asr_realtime/conf/conf.ini

参数名	描述
process_host_list	识别层服务对应的 ip 与 port 列表
cgi_port	接入层服务对应的 port

process_port	识别层服务对应的 port
mode_type	识别模型类型 16k，8k
hotword_id	是否开启热词标识，当设置为 1 时，为开启热词，0 为关闭

运行./admin/中的 chage.sh 脚本将相关配置分发到服务正确位置

sh chage.sh

注意：

第一次部署或修改配置以后，一定要运行 chage.sh，否则无法将配置同步到各个模块，造成启动失败

3.4 启动服务（单机部署）

运行./admin/中的总控脚本 start.sh 完成服务启动

sh start.sh

```
[root@TENCENT64 /data/home/liqiansun/asr_realtime/admin]# sh start.sh
rec_process_home: /data/home/liqiansun/asr_realtime/process
[2020年 02月 11日 星期二 20:39:05 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:10 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:15 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:20 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:25 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:30 CST][wait to be ready][status:1]
[2020年 02月 11日 星期二 20:39:35 CST][wait to be ready][status:1]
aai_recognize start succeed
rec_cgi_home: /data/home/liqiansun/asr_realtime/cgi
rec_cgi_bin: /data/home/liqiansun/asr_realtime/cgi/bin/asr_api_server
service_conf: /data/home/liqiansun/asr_realtime/cgi/bin/conf/online/asr_api_server.conf
log_name: .nohup.log.2020021120
server port: 9090
status: 1
[2020年 02月 11日 星期二 20:39:40 CST][wait to be ready][status:1]
status: 0
asr_api_server start succeed
Create crontab task...
Create crontab task complete.
```

启动成功脚本退出，可以查看进程再次确认服务已经正常启动：

ps aux | grep -E "aai_recognize|asr_api_server”

或查看端口是否正常启动：

netstat -anp | grep -E "9090|7800"

```
tcp        0      0 0.0.0.0:7800            0.0.0.0:*               LISTEN      51714/aai_recognize
tcp6       0      0 :::9090                 :::*                    LISTEN      14947/asr_api_serve
```

停止/重启总控脚本：stop.sh/restart.sh

注意：

总控脚本会将服务注册到守护进程中，保证服务异常退出后自动启动，因此要正确使用 stop.sh 停止服务

3.5 多机部署和集群部署

步骤一：替换授权和修改配置文件，同 3.2、3.3，需要说明集群部署时，process_host_list，将其修改为对应的识别层机器 ip 与端口，若有多台识别层机器，用逗号分隔，示例如图所示

```
process_host_list=["127.0.0.1:7800","10.148.173.32:7800"]
cgi_port=9090
process_port=7800
```

步骤二：分别运行./admin/中的 chage.sh 脚本将相关配置分发到服务正确位置

sh chage.sh

步骤三：分别启动接入层、识别层

启动接入层，执行单独启动 admin 下 cgi 层脚本：

sh start_cgi.sh

启动识别层，执行单独启动 admin 下识别层脚本：

sh start_process.sh

3.6 服务验证

（1）测试用例

当前提供了基于 php/python 开发的 demo，测试用例存放路径：

asr_realtime/test/php

asr_realtime/test/python

以 python 为例，修改主要参数： test/python/ Config.py

```
class Config:
    '全局变量配置信息，请按需求改成自己的配置'

    # ------------- optional, 根据自身需求配置值 ---------------
    servicepath= "http://127.0.0.1:9090/realtime_asr_private"
    filepath = "../testvoice/8k.wav"
    # 结果返回方式 0: 同步返回 or 1: 尾包返回
    RES_TYPE = 0
    # 识别结果文本编码方式 0:UTF-8, 1:GB2312, 2:GBK, 3:BIG5
    RESULT_TEXT_FORMAT = 0
    #  语音编码方式 1:wav  4:sp  6:skill 8:mp3
    VOICE_FORMAT = 1
    # 语音切片长度 cutlength<200000
    CUT_LENGTH = 8192

config = Config()
```

运行：*python test/python/RasrClient.py*

如果有识别结果，说明服务部署正常。若报错，参考下面常见错误码和日志信息

（2）错误码说明

返回码	说明
100	获取语音分片信息失败

101	语音分片过大
102	参数不合法
110	后台识别服务器故障，请从 seq=0 重传
111	后台识别模块回包格式错误
112	语音分片为空
113	后台服务器识别超时
115	时长计算时音频类型不合法
116	无授权权限
120	获取 rpcClient 错误
121	后台识别服务器错误，请从 seq=0 重传
122	后台识别服务器收到的包格式错误
123	后台识别服务器音频解压失败，请从 seq=0 重传
124	后台识别服务器识别失败，请从 seq=0 重传
125	后台识别服务器识别失败，请重新尝试
126	后台识别服务器音频分片等待超时，请从 seq=0 重传
127	后台识别服务器音频分片重复

（3）日志查看

接入层日志路径： asr_realtime/cgi/log

```
[root@TENCENT64 /data/home/liqiansun/asr_realtime/cgi/log]# ll
总用量 220
-rw-rw-r-- 1 root root   8397 2月  24 16:16 asr_api_server.log
-rw-r----- 1 root root   6292 2月  22 02:09 asr_api_server.log.2020022202
-rw-r----- 1 root root   6259 2月  22 03:11 asr_api_server.log.2020022203
-rw-rw-r-- 1 root root   1756 2月  24 11:51 asr_api_server.log.wf
-rw-r----- 1 root root   1757 2月  22 02:09 asr_api_server.log.wf.2020022202
-rw-r----- 1 root root   1748 2月  22 03:11 asr_api_server.log.wf.2020022203
-rw-rw-r-- 1 root root 181529 2月  24 16:16 asr_api_server_rt_dump.log
-rw-rw-r-- 1 root root      0 2月  24 16:13 asr_api_server_rt_dump.log.wf
```

识别层日志路径：asr_realtime/process/log。比如看查看错误信息:查看 asr_realtime.ERROR

```
[root@TENCENT64 /data/home/liqiansun/asr_realtime/process/log]# ll
总用量 568
lrwxrwxrwx 1 root root     64 2月  24 16:16 asr_realtime.ERROR -> asr_realtime.TENCENT64.site.root.log.ERROR.20200224-161643.
12104
lrwxrwxrwx 1 root root     63 2月  24 16:12 asr_realtime.INFO -> asr_realtime.TENCENT64.site.root.log.INFO.20200224-161227.12
104
-rw-r--r-- 1 root root    612 2月  24 11:58 asr_realtime.TENCENT64.site.root.log.ERROR.20200224-115803.10976
-rw-r--r-- 1 root root   3152 2月  24 16:16 asr_realtime.TENCENT64.site.root.log.ERROR.20200224-161643.12104
-rw-r--r-- 1 root root   1370 2月  24 11:58 asr_realtime.TENCENT64.site.root.log.INFO.20200224-115759.10976
-rw-r--r-- 1 root root 540989 2月  25 12:02 asr_realtime.TENCENT64.site.root.log.INFO.20200224-161227.12104
-rw-r--r-- 1 root root    612 2月  24 11:58 asr_realtime.TENCENT64.site.root.log.WARNING.20200224-115803.10976
-rw-r--r-- 1 root root   3372 2月  24 16:16 asr_realtime.TENCENT64.site.root.log.WARNING.20200224-161643.12104
lrwxrwxrwx 1 root root     66 2月  24 16:16 asr_realtime.WARNING -> asr_realtime.TENCENT64.site.root.log.WARNING.20200224-161
643.12104
```

更多运维相关内容，请参考《腾讯云实时识别运维文档》。

返回顶部