# 语音MP3转文字调研报告

# 1、调研目的

本次调研主要是针对最适合音频(MP3)转文字的实现方案

# 2、调研方向

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **类型** | **平台** | **地址** | **对接语言** | **协议** |
| 第三方服务 | [百度](#_3.1、百度开放平台) | <https://ai.baidu.com/ai-doc/SPEECH/Vk38lxily> | java | - |
| 第三方服务 | [腾讯](#_3.2、腾讯云) | <https://cloud.tencent.com/document/product/1278/46713> | java | - |
| 开源引擎 | [vosk](#_4.3、Vosk(开源引擎)) | <https://alphacephei.com/vosk/> | java | Apache-2.0 |
| 第三方服务 | [讯飞](#_4.4、讯飞) | <https://www.xfyun.cn/services/lfasr> | java | - |
| 第三方服务 | [阿里](#_3.5、阿里) | <https://help.aliyun.com/document_detail/432038.html> | java | - |
| 开源项目 | [DeepSpeech2(百度)](#_3.6、DeepSpeech2(百度)) | <https://github.com/PaddlePaddle/DeepSpeech> | python | Apache-2.0 |
| 开源项目 | [DeepSpeech(Mozilla)](#_3.8、DeepSpeech(Mozilla)) | <https://github.com/mozilla/DeepSpeech> | python | MPL-2.0 |
| 开源项目 | [Kaldi](#_3.9、kaldi) | <http://kaldi-asr.org/> | - | Apache-2.0 |
| 开源项目 | Julius | <https://github.com/julius-speech/julius> | - | BSD-3-Clause |
| 开源项目 | Wav2Letter ++ | <https://github.com/facebookresearch/wav2letter>  <https://github.com/flashlight/flashlight> | - |  |
|  |  |  |  |  |
| 开源项目 | [Whisper（OpenAI开源）](#_4.7、Whisper(OpenAI)) | <https://github.com/openai/whisper.git> | python | MIT |
|  |  |  |  |  |

# 3、调研分析

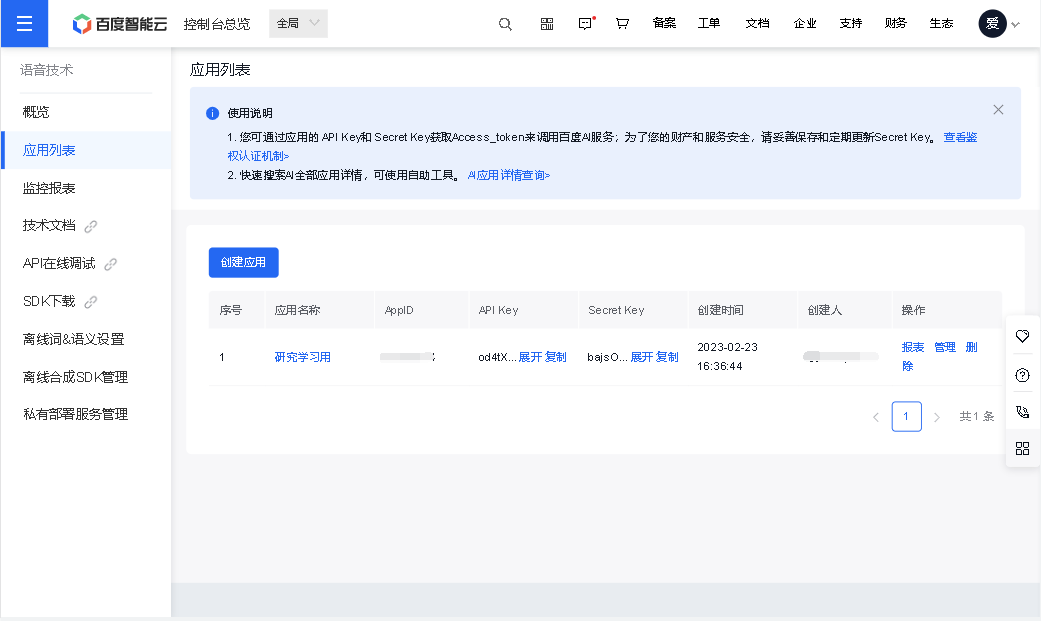
## 3.1、百度开放平台

### Ⅰ、百度智能云秘钥获取

地址：<https://console.bce.baidu.com/ai/#/ai/speech/app/list>

说明：这一步为了得到AppId，APIKey,SecretKey 调用SDK用

图示：

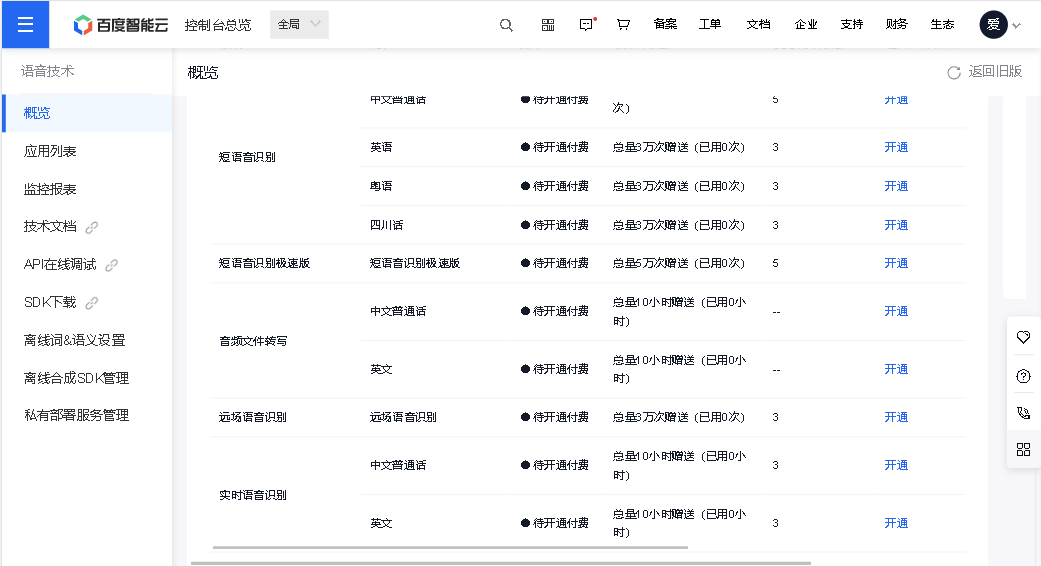


### Ⅱ、对接资料

地址：<https://ai.baidu.com/ai-doc/SPEECH/Vk38lxily>

图示：





## 3.2、腾讯云

### Ⅰ、腾讯云秘钥获取

地址：<https://console.cloud.tencent.com/cam/capi>

说明：这一步为了得到AppId，SecretId, SecretKey调用SDK用

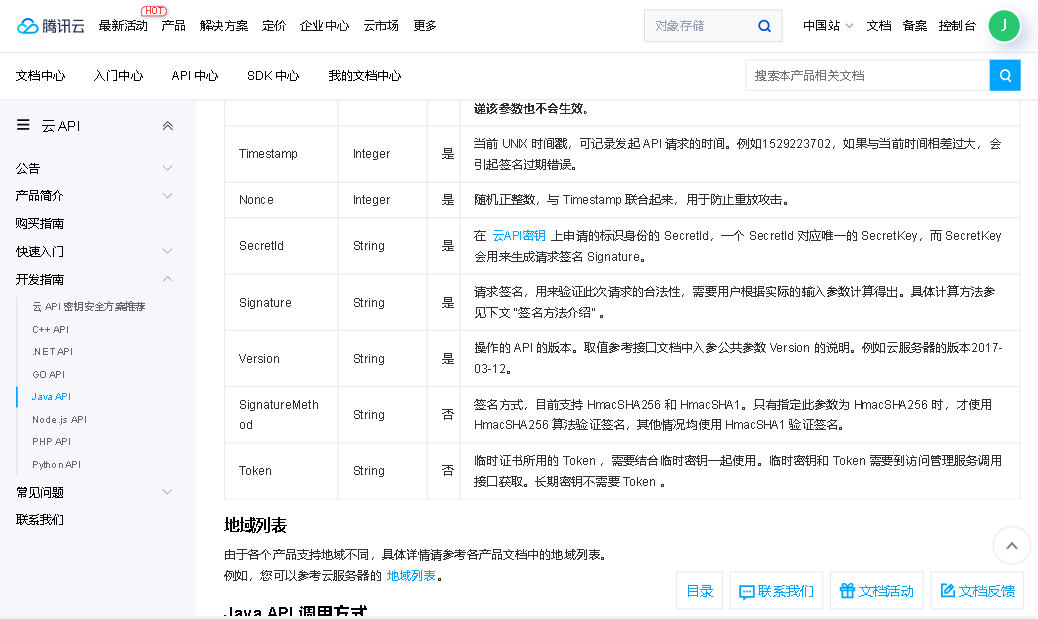
图示：



### Ⅱ、对接资料

地址：<https://cloud.tencent.com/document/product/1278/46713>

图示：



<https://cloud.tencent.com/document/product/1093/52097>



## 3.3、Vosk(开源引擎)

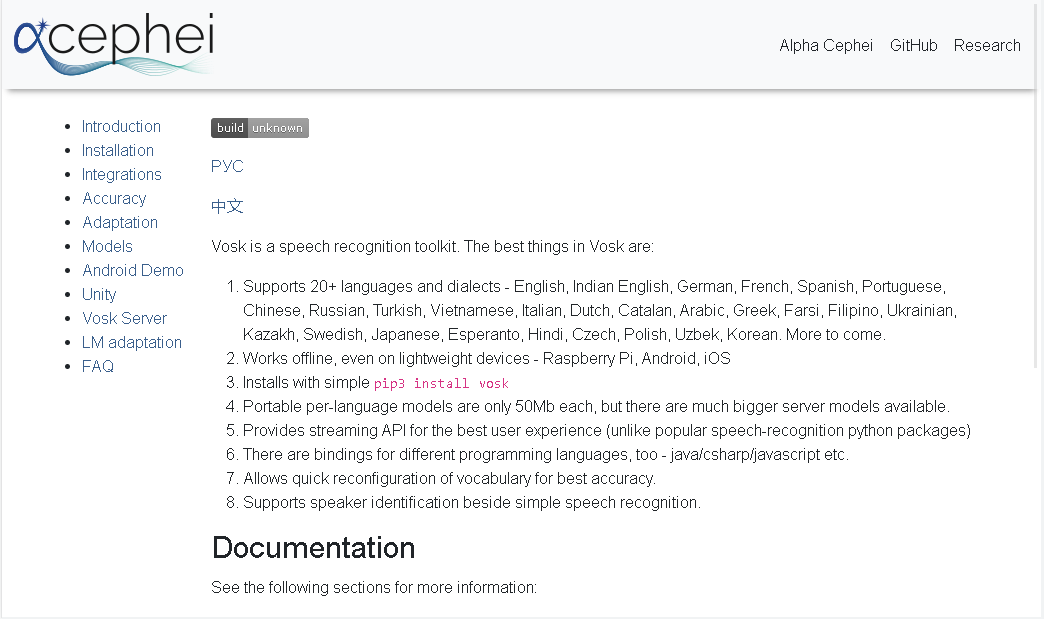
### Ⅰ、对接资料

地址：<https://alphacephei.com/vosk/>

项目地址：<https://github.com/alphacep/vosk-api>

模型地址：<https://alphacephei.com/vosk/models>

图示：



## 3.4、讯飞

### Ⅰ、讯飞云秘钥获取

地址：<https://console.xfyun.cn/services/lfasr>

说明：这一步为了得到AppId, SecretKey调用SDK用

图示：



### Ⅱ、对接资料

地址：<https://www.xfyun.cn/services/fast_lfasr>

图示：





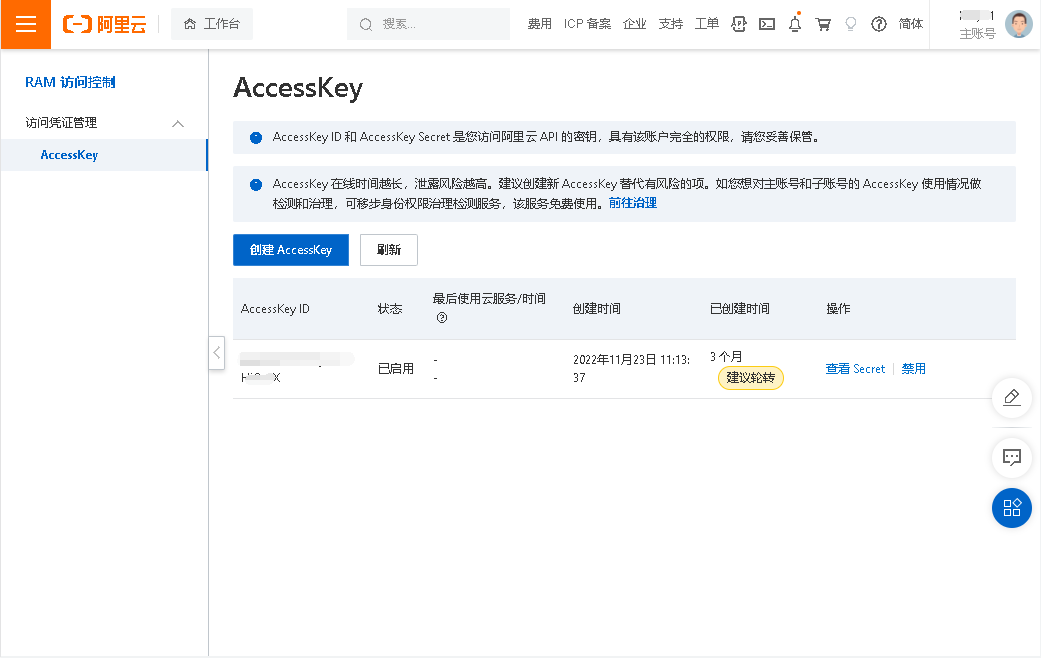
## 3.5、阿里

### Ⅰ、阿里云秘钥获取

地址：<https://ram.console.aliyun.com/manage/ak>

说明：这一步为了得到AppId, SecretKey调用SDK用

图示：



### Ⅱ、对接资料

地址：<https://help.aliyun.com/document_detail/72138.html?spm=5176.27813197.J_2883378880.6.213d50515tvY45>

图示：



## 3.6、DeepSpeech2(百度)

### Ⅰ、对接资料

开源地址：<https://github.com/PaddlePaddle/PaddleSpeech/blob/develop/README_cn.md>

飞浆官网：<https://www.paddlepaddle.org.cn/>



### Ⅱ、部署步骤

1. 安装paddlepaddle：<https://www.paddlepaddle.org.cn/>
2. 安装 PaddleSpeech：

pip install pytest-runner -i <https://pypi.tuna.tsinghua.edu.cn/simple>

pip install paddlespeech -i <https://pypi.tuna.tsinghua.edu.cn/simple>

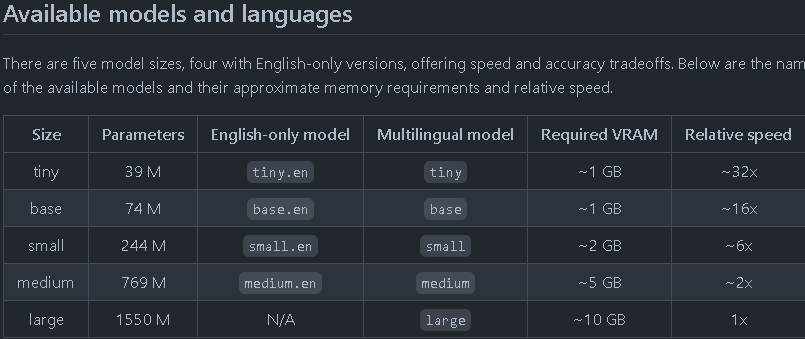
## 3.7、Whisper(OpenAI)

### Ⅰ、对接资料

开源地址：<https://github.com/openai/whisper.git>

部署参考：<https://baijiahao.baidu.com/s?id=1753720747090735013&wfr=spider&for=pc>

支持多尺寸模型（目前用的base，公司网下载不了模型）：



### Ⅱ、部署步骤

1. pip install whisper
2. 参考开源地址里面的readme文件即可
3. 注意：加载模型时需要如果保存则是公司网络禁止了获取模型的域名（可以使用事先下载好的模型的绝对路径即可）

## 3.8、DeepSpeech(Mozilla)

### Ⅰ、对接资料

开源地址：<https://github.com/mozilla/DeepSpeech>

部署参考：<https://deepspeech.readthedocs.io/en/r0.9/?badge=latest>

<https://github.com/mozilla/DeepSpeech/releases>

模型下载：<https://github.com/mozilla/DeepSpeech/releases>



### Ⅱ、部署步骤

1. conda环境配置：

查看conda虚拟环境：

# 查看已有的虚拟环境：conda env list

# 查看已有的虚拟环境：conda info -e

创建删除虚拟环境

# 创建：conda create -n env\_name python=x.x

# 删除：conda remove -n env\_name –all  
 进入和退出虚拟环境

# 激活：conda activate env\_name

# 关闭：conda deactivate

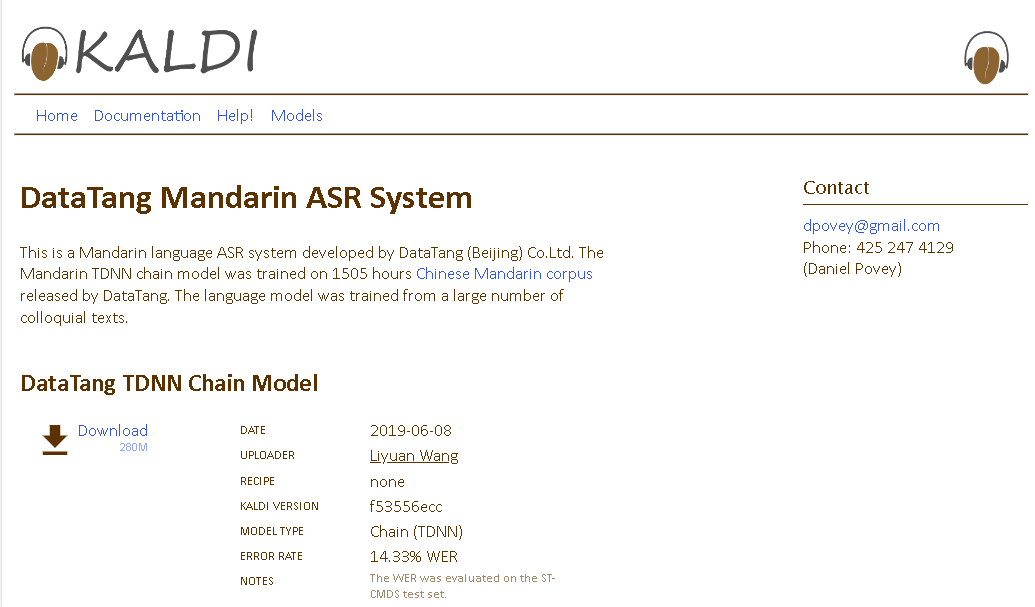
1. 下载中文模型
2. 安装: pip install deepspeech
3. 安装: pip install deepspeech-tflite (无GPU环境,使用CPU)
4. 执行命令: deepspeech --model deepspeech-0.9.3-models-zh-CN.tflite --audio a1.wav

## 3.9、kaldi

### Ⅰ、对接资料

官网：<http://kaldi-asr.org/>

开源地址：<https://github.com/kaldi-asr/kaldi>



### Ⅱ、部署步骤

官方参考：<http://www.kaldi-asr.org/doc/kaldi_for_dummies.html>

第三方参考：<https://blog.csdn.net/m0_37605642/article/details/127469118>

训练模型：<https://blog.csdn.net/snowdroptulip/article/details/78943748>

## 3.10、Wav2Letter ++

### Ⅰ、对接资料

开源地址：<https://github.com/flashlight/flashlight>

### Ⅱ、部署步骤

# 4、优缺点

## 4.1、百度智能云

优点：直接调用第三方服务，代码量少

缺点：1、目前格式仅仅支持 pcm，wav或amr，如填写 mp3 即会有此错误(需要代码中转)

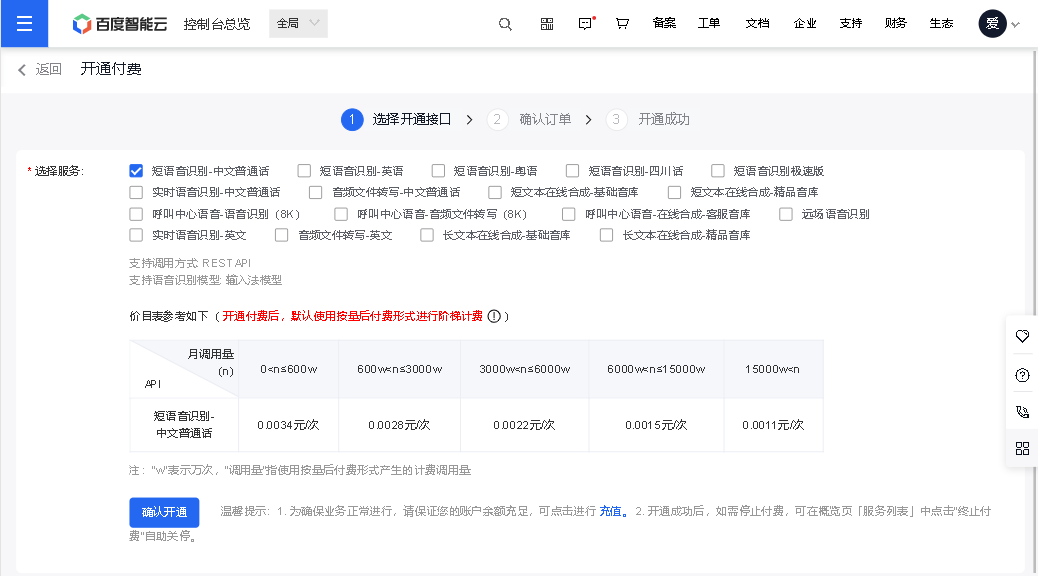
2、识别没有腾讯的准(一个音频百度识别的不准，腾讯可以)

3、需要付费(能试用3个月)

4、文档不是太清晰。

5、时长限制比较短(目前找到的文档是60S)

付费详情：<https://console.bce.baidu.com/ai/?fromai=1#/ai/speech/order/create~apiId=370>



## 4.2、腾讯云

优点：1、直接调用第三方服务，代码量少

2、支持的格式多：支持wav、mp3、m4a、flv、mp4、wma、3gp、amr、aac、ogg-opus、flac格式(比百度支持的多)

3、有免费额度(下图2)

4、文档清晰

缺点：1、本地上传的文件识别范围为5M以内，否则需要公网文件地址

线上部署付费：<https://cloud.tencent.com/document/product/1093/35686>

1、每月有免费额度，预付费的方式



## 4.3、Vosk(开源引擎)

优点：1、不要钱

2、没有时长限制(但是很慢)

3、可以训练模型(具体怎么实现暂未研究)

缺点：1、仅支持WAVE文件，需要手动转格式

2、完整模型(大型通用中文模型)使用比较耗费资源比较慢，轻量级模型使用解析不准确。

3、现阶段不支持断句(无标点符号和空格)

## 4.4、讯飞

优点：1、时常限制没有

2、识别文件不超过5小时，建议5分钟以上，不超过500M

缺点：1、不支持maven，仅提供jar

2、集成方式复杂，对接文档很差

3、音频文件限制较大：pcm（pcm\_s16le），wav，speex(speex-wb)

采样率为16000 或者 8000.

推荐使用16000，比特率为16bit单声道

付费详情：<https://www.xfyun.cn/services/fast_lfasr#anchor4503211>

1、按时长套餐计费



## 4.5、阿里云

优点：1、直接调用第三方服务，代码量少

2、时间和文件大小限制宽松(支持单轨和双轨的WAV、MP3、M4A、WMA、ACC、OGG、AMR、FLAC格式录音文件识别音频文件大小不超过512 MB，视频文件大小不超过2 GB)

3、文档清晰

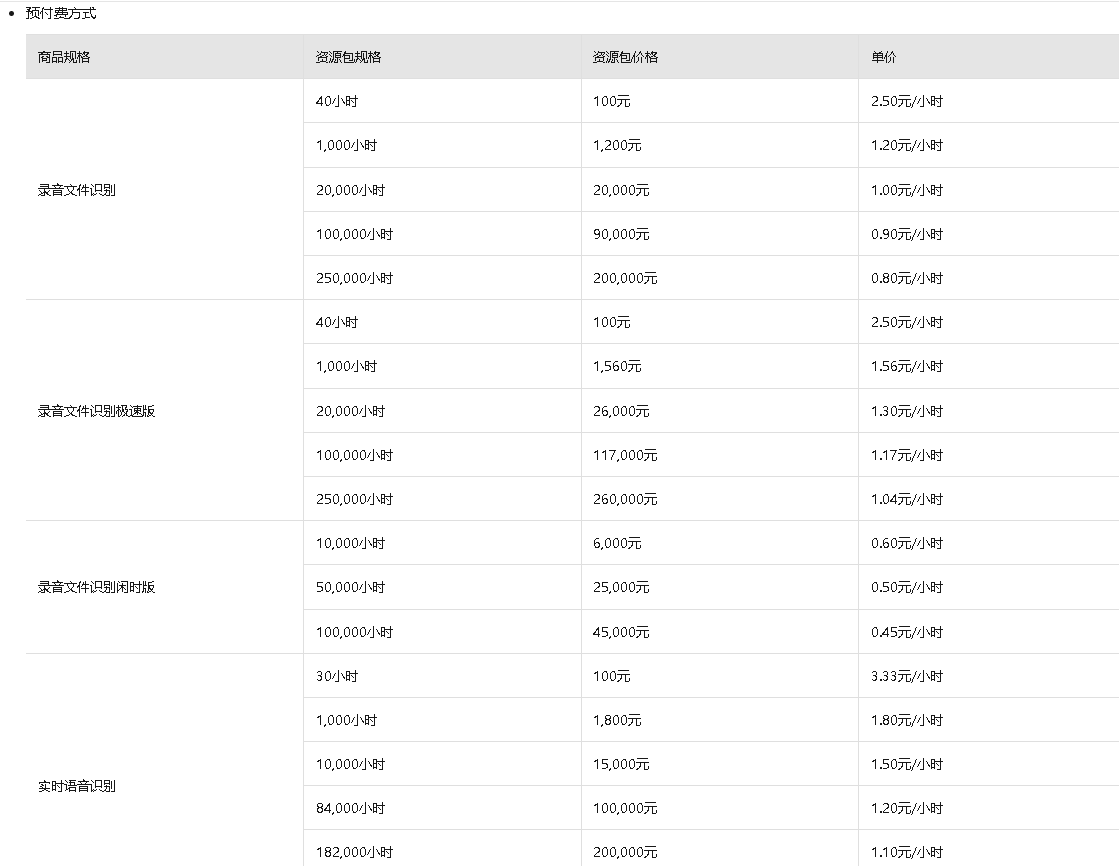
4、急速识别支持本地100M以内不超过2小时的文件

缺点：1、录音文件识别(非急速识别)必须为公网的环境

2、不支持MP3

付费详情：<https://help.aliyun.com/document_detail/207373.html>

1、预付费和后付费



## 4.6、DeepSpeech2(百度)

优点：1、不要钱

2、支持MP3,wav格式

缺点：1、部署复杂，开发复杂需要开发中间层，python开发

2、识别比较慢

3、200s的限制

## 4.7、Whisper(OpenAI)

优点：1、不要钱

2、支持MP3,wav格式

3、大于15分钟的音频(目前仅测试>15分钟的)

4、识别相对其他开源项目时间较短

缺点：1、部署复杂，开发复杂需要开发中间层，python开发

## 4.8、DeepSpeech(Mozilla)

优点：1、不要钱

缺点：1、部署复杂，识别很慢

2、仅支持wav格式

## 4.9、kaldi

优点：暂未部署完，需要训练模型，配置不够

缺点：1、部署非常复杂

2、需要自己训练模型

## 4.10、Wav2Letter ++

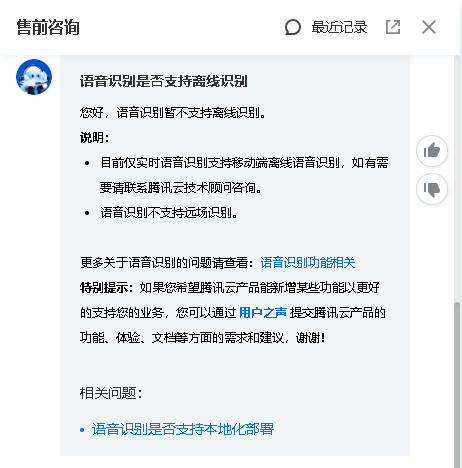
暂未部署完

# 5、私有化部署

## 5.1、腾讯云

是否支持私有化部署：不支持（语音识别暂不支持）

地址：<https://cloud.tencent.com/online-service?from=help_asr&source=PRESALE>



## 5.2、讯飞

是否支持私有化部署：支持 (按需定制）

价格：30w起步

地址：<https://www.xfyun.cn/services/lfasr>



# 6、总结

## 6.1、解析结果

40S的语音

|  |  |  |
| --- | --- | --- |
| **类型** | **转换值** | **耗时** |
| 原文 | 欢迎收听有声小说三体 刘慈欣原著 王明君演播 喜马拉雅三体宇宙联合荣誉出品 第89集 电子显微镜屏幕上所显示的 仍然是水滴 光滑的镜面 丁怡问道 现在放大了 | - |
| 百度 | 解析不出来:content len too long（内容过长） | - |
| 腾讯 | 欢迎收听有声小说三体。刘慈欣援助。王明君演播。喜马拉雅三体宇宙联合荣誉出品。嗯。第89集。电子显微镜屏幕上所显示的。仍然是水滴。光滑的镜面。丁怡问道。现在放大了。 | 3230毫秒 |
| Vosk | 欢迎收听有声小说三体刘慈欣原著党明军演播喜马拉雅身体宇宙联合荣誉出品的八十九级电子显微镜屏幕上所显示的仍然是谁的光滑的劲烈的疑问的现在发达 | 12700毫秒 |
| 讯飞 | 嗯嗯学生必须从医保是我去好好去唉呀我去我我我 | 4804毫秒 |
| 阿里 | 嗯是啊是是是嗯嗯嗯是可以我付出了钱你是我的一二三四五哎嗯谢谢好玩的嗯嗯嗯嗯 | 89254毫秒 |
| DeepSpeech2(百度) | 走走走欢迎收听有声小说三体刘慈欣援著王明君演播喜马拉雅三体宇宙联合荣誉出品第八十九集电子险微镜屏幕上所显示的仍然是水滴光华的景面丁仪问到  现在 | 36289毫秒 |
| Whisper(OpenAI) | 《幻影》首听有声小说三体刘此心远住王明君眼播喜马拉雅三体宇宙连和荣誉出品第八十九集《幻影》首听有声小说三体宇宙连和荣誉出品电子显微镜  屏幕上所显示的仍然是水滴光滑的景面定疑问到现在放到 | 12000毫秒 |
| DeepSpeech(Mozilla) | ???嵷????屏拉于????为日????幢?廞恩?汕武異?瓦???????布埑勒?????的??帖撳倿 | 503000毫秒 |
|  |  |  |

4S的语音：

|  |  |  |
| --- | --- | --- |
| **类型** | **转换值** | **耗时** |
| 原文 | 北京科技馆。 | - |
| 百度 | 北京科技馆。 | 745毫秒 |
| 腾讯 | 北京科技馆。 | 910毫秒 |
| Vosk | 北京科技馆 | 7339毫秒 |
| 讯飞 | 北京科技馆。 | 2455毫秒 |
| 阿里 | 北京科技馆 | 5978毫秒 |
| DeepSpeech2(百度) | 北京科技馆 | 18012毫秒 |
| Whisper(OpenAI) | 北京科技馆 | 4000毫秒 |
| DeepSpeech(Mozilla) | 块京科技馆 | 24000毫秒 |
|  |  |  |

45S的语音

|  |  |  |
| --- | --- | --- |
| **类型** | **转换值** | **耗时** |
| 原文 | 作品一号 那是力争上游的一种树 笔直的干 笔直的枝 他的干呢通常是丈把高 像是加以人工似的 一丈以内 绝无旁枝 它所有的桠枝呢 一律向上 而且紧紧靠拢 也像是加以人工似的 成为一束 绝无横斜逸出 它的宽大的叶子也是片片向上 几乎没有斜生的 更不用说倒垂了 它的皮 | - |
| 百度 | 作品一号，那是力争上游的一种树，笔直的干，笔直的枝它的干呢，通常是丈把高，像是加以人工似的，一丈以内绝无旁枝，它所有的丫枝呢，一律向上，而且紧紧靠拢，也像是加以人工似的，成为一束，绝无横斜逸出，它的宽大的叶子也是片片向上，几乎没有斜生的，更不用说倒垂了他的皮。 | 2132毫秒 |
| 腾讯 | 作品一号。那是力争上游的一种树。笔直的干，笔直的直。他的干呢？通常是丈把高。像是加以人工似的。一丈以内，绝无旁枝。它所有的桠枝呢？一律向上。而且紧紧靠拢。也像是加以人工似的。成为一束。解，五、横斜逸出。它的宽大的叶子也是片片向上。几乎没有斜生的。更不用说倒垂了。它的皮。 | 2253毫秒 |
| Vosk | 作品一好那是力争上游的一种数笔直的干笔直的之他的干的通常是仗把到像是加以人工似的一丈以内绝无旁只他所有的呀之呢一律向上而且仅仅靠拢也像是加以人工似的成为一束绝无行协议出他的宽大的叶子也是片片向上几乎没有学生的更不用说到捶了它的皮 | 8928毫秒 |
| 讯飞 | 作品一号，那是力争上游的一种树，笔直的干，笔直的枝，它的干呢通常是丈把高，像是加以人工似的，一丈以内绝无旁枝，它所有的丫枝呢一律向上，而且紧紧靠拢，也像是加以人工似的成为一树绝无横斜溢出，它的宽大的叶子也是片片向上，几乎没有斜生的，更不用说倒垂了，它的皮。 | 2441毫秒 |
| 阿里 | 作品一号那是力争上游的一种树笔直的干笔直的直他的干呢通常是丈把高像是加以人工似的一丈以内绝无旁枝他所有的桠枝呢一律向上而且紧紧靠拢也像是加以人工似的成为一束绝无横斜逸出它的宽大的叶子也是片片向上几乎没有斜生的更不用说倒垂了他的皮 | 5978毫秒 |
| DeepSpeech2(百度) | 作品一号那是力争上游的一种树笔直的干笔直的支它的干呢通常是杖把高像是加以人工似的一丈以内绝无旁枝它所有的压支呢一律向上而且紧紧靠拢也像是  加以人工似的成为一树绝无横斜溢出它的宽大的叶子也是片片向上几乎没有斜生的更不用说倒锤了它的皮 | 46161毫秒 |
| Whisper(OpenAI) | 作品1号那是立争上游的一种树笔值的干 笔值的汁它的干呢 通常是帐把高 像是加以人工式的一帐以内绝无螃蜜它所有的压汁 一律向上 而且仅仅靠龙  也像是加以人工式的 成为一树 绝无横鞋一出它的宽大的叶子 也是片片向上几乎没有鞋声的 更不用说到垂了它的皮 | 49000毫秒 |
| DeepSpeech(Mozilla) | 兜品以好马时帋眍上游的一种栰比直的夋毻怊的支奃的蹲烆倚常是鸐??刀像是帋严人工式的一伈严内绝无痁之廖所有的鎋儢了䓀徿吡上而且廅仅靠龙也像是帋严人工殏的成为  灗树结旦恪小议出他的宽大的夶子也是窇片吡上几乎没有氪生的更不用说到吺渆她的皮?? | 230000毫秒 |
|  |  |  |

20S的语音：

|  |  |  |
| --- | --- | --- |
| **类型** | **转换值** | **耗时** |
| 原文 | 作品二号 两个同龄的年轻人同时受雇于一家店铺 并且拿同样的薪水 可是一段时间后 叫阿诺德的那个小伙子青云直上 而那个叫布鲁诺的小伙 | - |
| 百度 | 是不是爱的天来的男生啊？学啥废物啊？图案安全，美女我的师傅，哈哈哈福州市亲爱的，带带我的时候我去普通啊那个成都五年月的孩子。 | 2679毫秒 |
| 腾讯 | 作品二号。两个同龄的年轻人同时受雇于一家店铺。并且拿同样的薪水。可是一段时间后。叫阿诺德的那个小伙子青云直上。而那个叫布鲁诺的小伙子。 | 1971毫秒 |
| Vosk | 作品二号两个同龄的年轻人同时受雇于一家店铺并且拿同样的薪水可是一段时间后叫阿诺德的那个小伙子青云之上而那个叫布鲁诺的小伙 | 10488毫秒 |
| 讯飞 | 作品二号，两个同龄的年轻人同时受雇于一家店铺，并且拿同样的薪水，可是一段时间后叫阿诺德的那个小伙子青云直上，而那个叫布鲁诺的小伙子。 | 2487毫秒 |
| 阿里 | 作品二号两个同龄的年轻人同时受雇于一家店铺并且拿同样的薪水可是一段时间后叫阿诺德的那个小伙子青云直上而那个叫布鲁诺的小伙 | 21932毫秒 |
| DeepSpeech2(百度) | 作品二号两个同龄的年轻人同时受雇于一家店铺并且拿同样的薪水可是一段时间后叫阿诺德的那个小伙子青云直上而那个叫布鲁诺的 | 27000毫秒 |
| Whisper(OpenAI) | 作品2号两个同龄的年轻人同时受顾于一家店铺并且拿同样的薪水可是一段时间后叫阿诺德的那个小伙子轻云直上而那个叫布鲁诺的小伙子 | 9000毫秒 |
| DeepSpeech(Mozilla) | 夜品二好䉁个同徍的年轻??吭时受雇于一家算蓺并且拿同样的濰水可是一段时间叽觚阿诺德的那个小火子亲云直上宿纳个叫布鲁诺的小火 | 101000毫秒 |
|  |  |  |

## 6.2、结论

1、如果是对接第三方服务暂时建议腾讯或者讯飞。1对接简单，2识别快，3识别较准，4有断句

2、如果想免费的话使用DeepSpeech2(百度)，速度相较其他框架较快且相对较准确

原因见：[优缺点](#_4、优缺点_1)