语音识别

API 文档

产品文档

腾讯云

文档目录

API 文档
实时语音识别相关接口
实时语音识别（websocket）

最近更新时间：2024-05-07 15:51:13

说明：
此接口为 API 2.0 版本，在参数风格、错误码等方面有区别于 API 3.0 版本，请知悉。

接口描述

本接口服务采用 websocket 协议，对实时音频流进行识别，同步返回识别结果，达到“边说边出文字”的效果。
在使用该接口前，需要在语音识别控制台开通服务，并进入 API 密钥管理页面 新建密钥，生成 AppID、SecretID 和 SecretKey，用于 API 调用时生成签名，签名将用来进行接口鉴权。

接口要求

集成实时语音识别 API 时，需按照以下要求。

内容	说明
语言种类	支持中文普通话、英语、韩语、日语、泰语、印度尼西亚语、越南语、马来语、菲律宾语、葡萄牙语、土耳其语、阿拉伯语、西班牙语、印地语、法语、德语。可通过接口参数 engine_model_type 设置对应语言类型
支持行业	通用、金融、游戏、教育、医疗
音频属性	采样率：16000Hz或8000Hz 采样精度：16bits 声道：单声道（mono）
音频格式	pcm、wav、opus、speex、silk、mp3、m4a、aac
请求协议	wss 协议
请求地址	wss://asr.cloud.tencent.com/asr/v2/<appid>?{请求参数}

接口鉴权	签名鉴权机制，详见 签名生成
响应格式	统一采用 JSON 格式
数据发送	建议每40ms 发送40ms 时长（即1:1实时率）的数据包，对应 pcm 大小为：8k 采样率640字节，16k 采样率1280字节 音频发送速率过快超过1:1实时率或者音频数据包之间发送间隔超过6秒，可能导致引擎出错，后台将返回错误并主动断开连接
并发限制	默认单账号限制并发连接数为50路，如您有提高并发限制的需求，请提工单 进行咨询

接口调用流程

接口调用流程分为两个阶段：握手阶段和识别阶段。两阶段后台均返回 text message，内容为 json 序列化字符串，以下是格式说明：

字段名	类型	描述
code	Integer	状态码，0代表正常，非0值表示发生错误
message	String	错误说明，发生错误时显示这个错误发生的具体原因，随着业务发展或体验优化，此文本可能会经常保持变更或更新
voice_id	String	音频流唯一 id，由客户端在握手阶段生成并赋值在调用参数中
message_id	String	本 message 唯一 id
result	Result	最新语音识别结果
final	Integer	该字段返回1时表示音频流全部识别结束

其中识别结果 Result 结构体格式为：

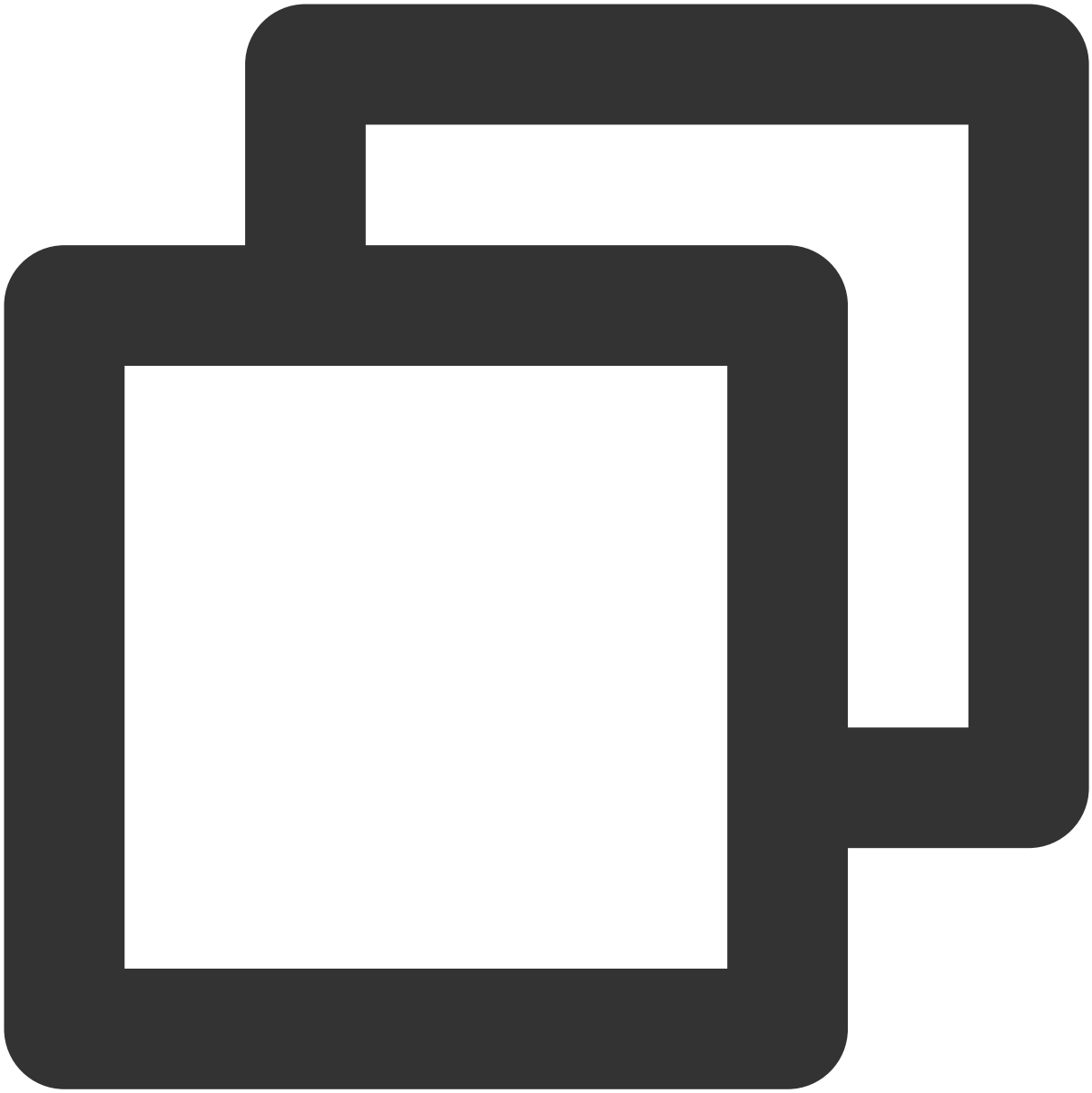
字段名	类型	描述
slice_type	Integer	识别结果类型： 0：一段话开始识别 1：一段话识别中，voice_text_str 为非稳态结果(该段识别结果还可能变化) 2：一段话识别结束，voice_text_str 为稳态结果(该段识别结果不再变化) 根据发送的音频情况，识别过程中可能返回的 slice_type 序列有： 0-1-2：一段话开始识别、识别中(可能有多次1返回)、识别结束

		0-2：一段话开始识别、识别结束 2：直接返回一段话完整的识别结果
index	Integer	当前一段话结果在整个音频流中的序号，从0开始逐句递增
start_time	Integer	当前一段话结果在整个音频流中的起始时间
end_time	Integer	当前一段话结果在整个音频流中的结束时间
voice_text_str	String	当前一段话文本结果，编码为 UTF8
word_size	Integer	当前一段话的词结果个数
word_list	Word Array	当前一段话的词列表，Word 结构体格式为： word：String 类型，该词的内容 start_time：Integer 类型，该词在整个音频流中的起始时间 end_time：Integer 类型，该词在整个音频流中的结束时间 stable_flag：Integer 类型，该词的稳态结果，0表示该词在后续识别中可能发生变化，1表示该词在后续识别过程中不会变化

握手阶段

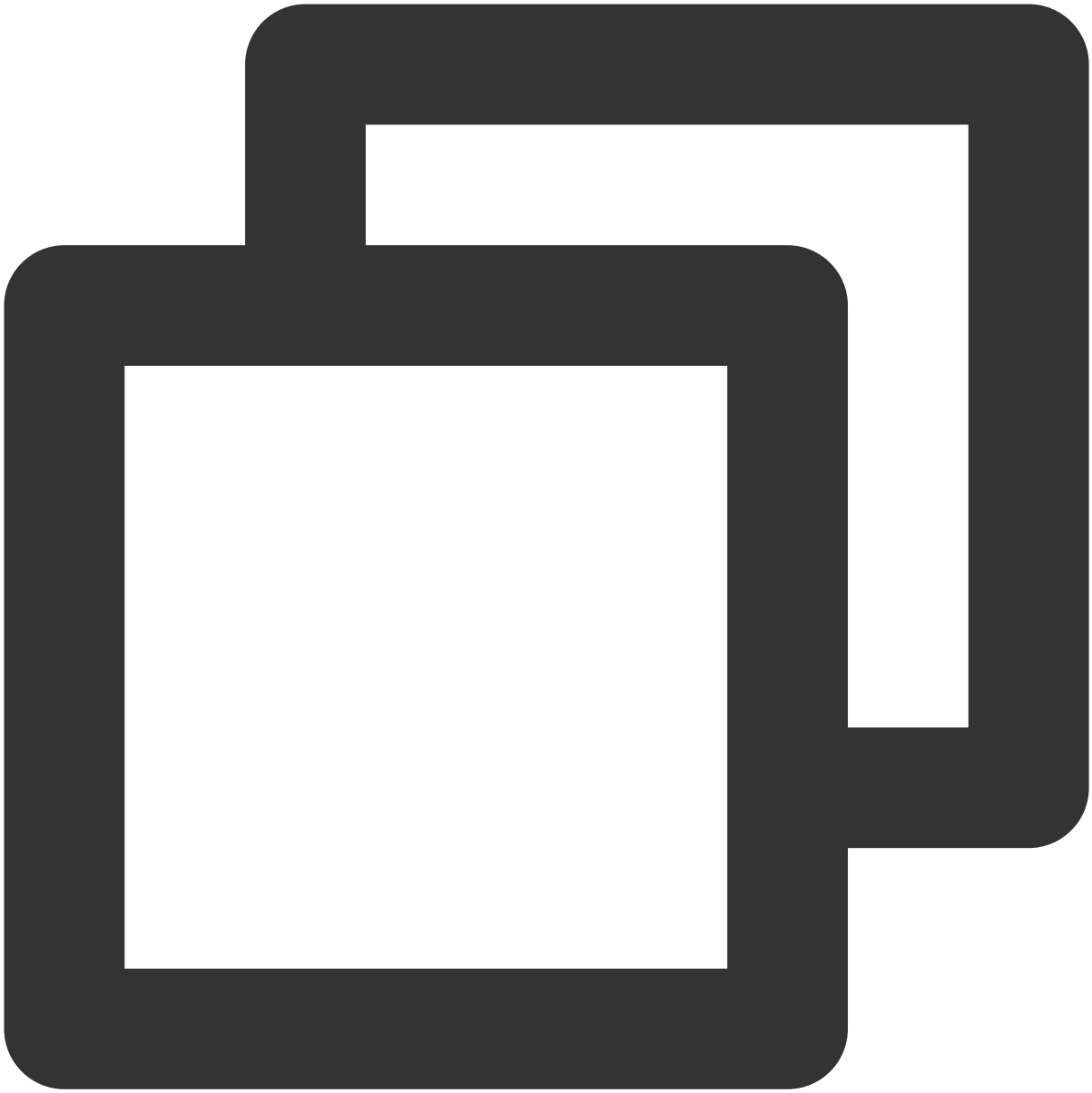
请求格式

握手阶段，客户端主动发起 websocket 连接请求，请求 URL 格式为：



```
wss://asr.cloud.tencent.com/asr/v2/<appid>?{请求参数}
```

其中<appid>需替换为腾讯云注册账号的 AppID，可通过 API 密钥管理页面 获取，{请求参数}格式为：



```
key1=value2&key2=value2...(key 和 value 都需要进行 urlencode)
```

参数说明：

参数名称	必填	类型	描述
secretid	是	String	腾讯云注册账号的密钥 SecretId，可通过 API 密钥管理页面 获取
timestamp	是	Integer	当前 UNIX 时间戳，单位为秒。如果与当前时间相差过大，会引起签名过期错误

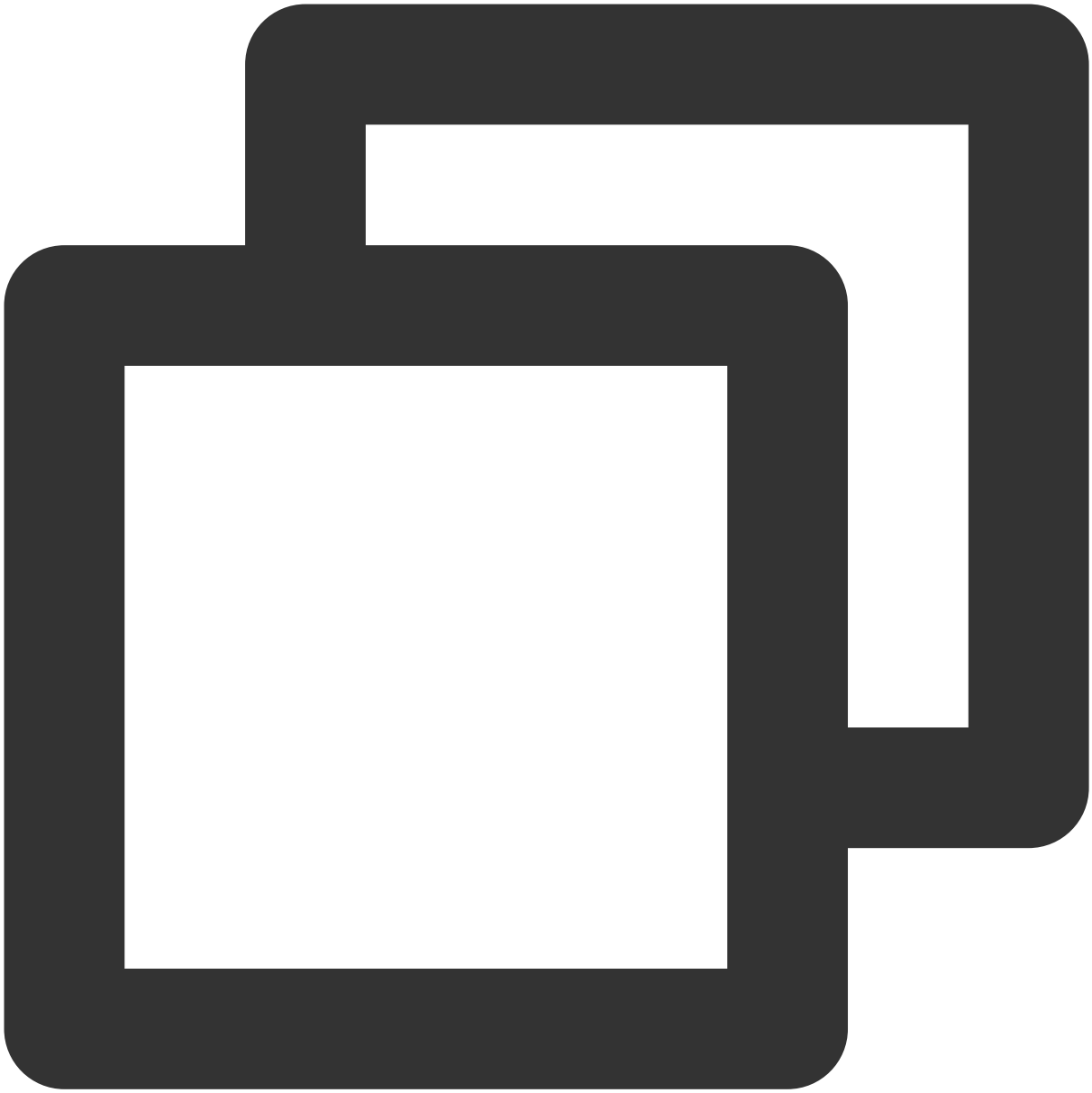
expired	是	Integer	签名的有效期截止时间 UNIX 时间戳，单位为秒。expired 必须大于 timestamp 且 expired - timestamp 小于90天
nonce	是	Integer	随机正整数。用户需自行生成，最长10位
engine_model_type	是	String	引擎模型类型 电话场景: • 8k_zh：中文电话通用； • 8k_en：英文电话通用； 非电话场景： • 16k_zh：中文普通话通用； • 16k_en：英语； • 16k_zh-PY：中英粤； • 16k_ko：韩语； • 16k_ja：日语； • 16k_th：泰语； • 16k_id : 印度尼西亚语； • 16k_vi: 越南语； • 16k_ms: 马来语； • 16k_fil: 菲律宾语； • 16k_pt：葡萄牙语； • 16k_tr：土耳其语； • 16k_ar: 阿拉伯语； • 16k_es: 西班牙语； • 16k_hi: 印地语； • 16k_fr: 法语； • 16k_de: 德语；
voice_id	是	String	16位 String 串作为每个音频的唯一标识，用户自己生成
voice_format	否	Int	语音编码方式，可选，默认值为4。1：pcm；4：speex(sp)；6：silk；8：mp3；10：opus（opus 格式音频流封装说明）；12：wav；14：m4a（每个分片须是一个完整的 m4a 音频）；16：aac
needvad	否	Integer	0：关闭 vad，1：开启 vad 如果语音分片长度超过60秒，用户需开启 vad（人声检测切分功能）
hotword_id	否	String	热词表 id。如不设置该参数，自动生效默认热词表；如果设置了该参数，那么将生效对应的热词表
reinforce_hotword	否	Integer	热词增强功能。默认为0，0：不开启，1：开启。 开启后（仅支持8k_zh，16k_zh），将开启同音替换功能，同音字、词在热词中配置。 举例：热词配置“蜜制”并开启增强功能后，与“蜜制”同拼音（mizhi）的“秘制”、“蜜汁”等的识别结果会被强制替换成“蜜

			制”。因此建议客户根据自己的实际情况开启该功能。
customization_id	否	String	自学习模型 id。如不设置该参数，自动生效最后一次上线的自学习模型；如果设置了该参数，那么将生效对应的自学习模型
filter_dirty	否	Integer	是否过滤脏词（目前支持中文普通话引擎）。默认为0。0：不过滤脏词；1：过滤脏词；2：将脏词替换为“ * ”
filter_modal	否	Integer	是否过语气词（目前支持中文普通话引擎）。默认为0。0：不过滤语气词；1：部分过滤；2：严格过滤
filter_punc	否	Integer	是否过滤句末的句号（目前支持中文普通话引擎）。默认为0。0：不过滤句末的句号；1：过滤句末的句号
filter_empty_result	否	Integer	是否回调识别空结果，默认为1。0：回调空结果；1：不回调空结果; **注意：**如果需要slice_type=0和slice_type=2配对回调，需要设置filter_empty_result=0。一般在外呼场景需要配对返回，通过slice_type=0来判断是否有人声出现。
convert_num_mode	否	Integer	是否进行阿拉伯数字智能转换（目前支持中文普通话引擎）。0：不转换，直接输出中文数字，1：根据场景智能转换为阿拉伯数字，3: 打开数学相关数字转换。默认值为1
word_info	否	Int	是否显示词级别时间戳。0：不显示；1：显示，不包含标点时间戳，2：显示，包含标点时间戳。支持引擎 8k_en、8k_zh、8k_zh_finance、16k_zh、16k_en、16k_ca、16k_zh-TW、16k_ja、16k_wuu-SH，默认为0
vad_silence_time	否	Integer	语音断句检测阈值，静音时长超过该阈值会被认为断句（多用在智能客服场景，需配合 needvad = 1 使用），取值范围：240-2000，单位 ms，此参数建议不要随意调整，可能会影响识别效果，目前仅支持 8k_zh、8k_zh_finance、16k_zh 引擎模型
max_speak_time	否	Integer	强制断句功能，取值范围 5000-90000（单位:毫秒），默认值0(不开启)。 在连续说话不间断情况下，该参数将实现强制断句（此时结果变成稳态，slice_type=2）。如：游戏解说场景，解说员持续不间断解说，无法断句的情况下，将此参数设置为10000，则将在每10秒收到 slice_type=2的回调。
noise_threshold	否	Float	噪音参数阈值，默认为0，取值范围：[-1,1]，对于一些音频片段，取值越大，判定为噪音情况越大。取值越小，判定为人声情况越大。 **慎用：可能影响识别效果**
signature	是	String	接口签名参数

hotword_list	否	String	临时热词表：该参数用于提升识别准确率。 单个热词限制："热词|权重"，单个热词不超过30个字符（最多10个汉字），权重1-11，如："腾讯云|5" 或 "ASR|11"； 临时热词表限制：多个热词用英文逗号分割，最多支持128个热词，如："腾讯云|10,语音识别|5,ASR|11"； 参数 hotword_id（热词表） 与 hotword_list（临时热词表） 区别： hotword_id：热词表。需要先在控制台或接口创建热词表，获得对应hotword_id传入参数来使用热词功能； hotword_list：临时热词表。每次请求时直接传入临时热词表来使用热词功能，云端不保留临时热词表。适用于有极大量热词需求的用户； 注意： 如果同时传入了 hotword_id 和 hotword_list，会优先使用 hotword_list； 热词权重设置为11时，当前热词将升级为超级热词，建议仅将重要且必须生效的热词设置到11，设置过多权重为11的热词将影响整体字准率。
input_sample_rate	否	Interge	支持 pcm 格式的8k音频在与引擎采样率不匹配的情况下升采样到16k后识别，能有效提升识别准确率。仅支持：8000。 如：传入 8000 ，则pcm音频采样率为8k，当引擎选用 16k_zh， 那么该8k采样率的 pcm 音频可以在16k_zh引擎下正常识别。 注：此参数仅适用于 pcm 格式音频，不传入值将维持默认状态，即默认调用的引擎采样率等于 pcm 音频采样率。

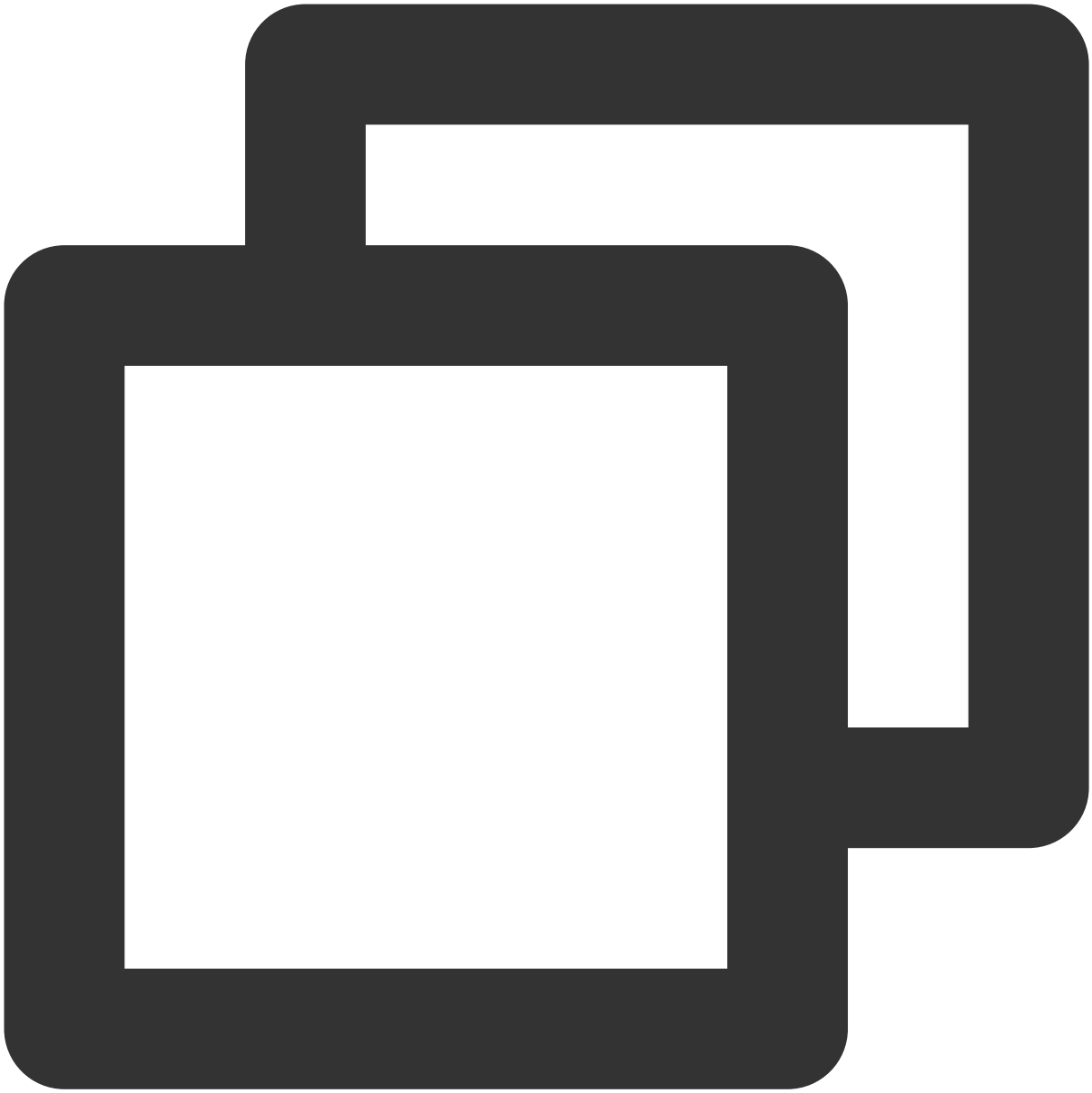
signature 签名生成

1. 对除 signature 之外的所有参数按字典序进行排序，拼接请求 URL 作为签名原文，这里以 Appid=1259228442, SecretId=AKIDoQq1zhZMN8dv0psmvud6OUKuGPO7pu0r 为例拼接签名原文，则拼接的签名原文为：



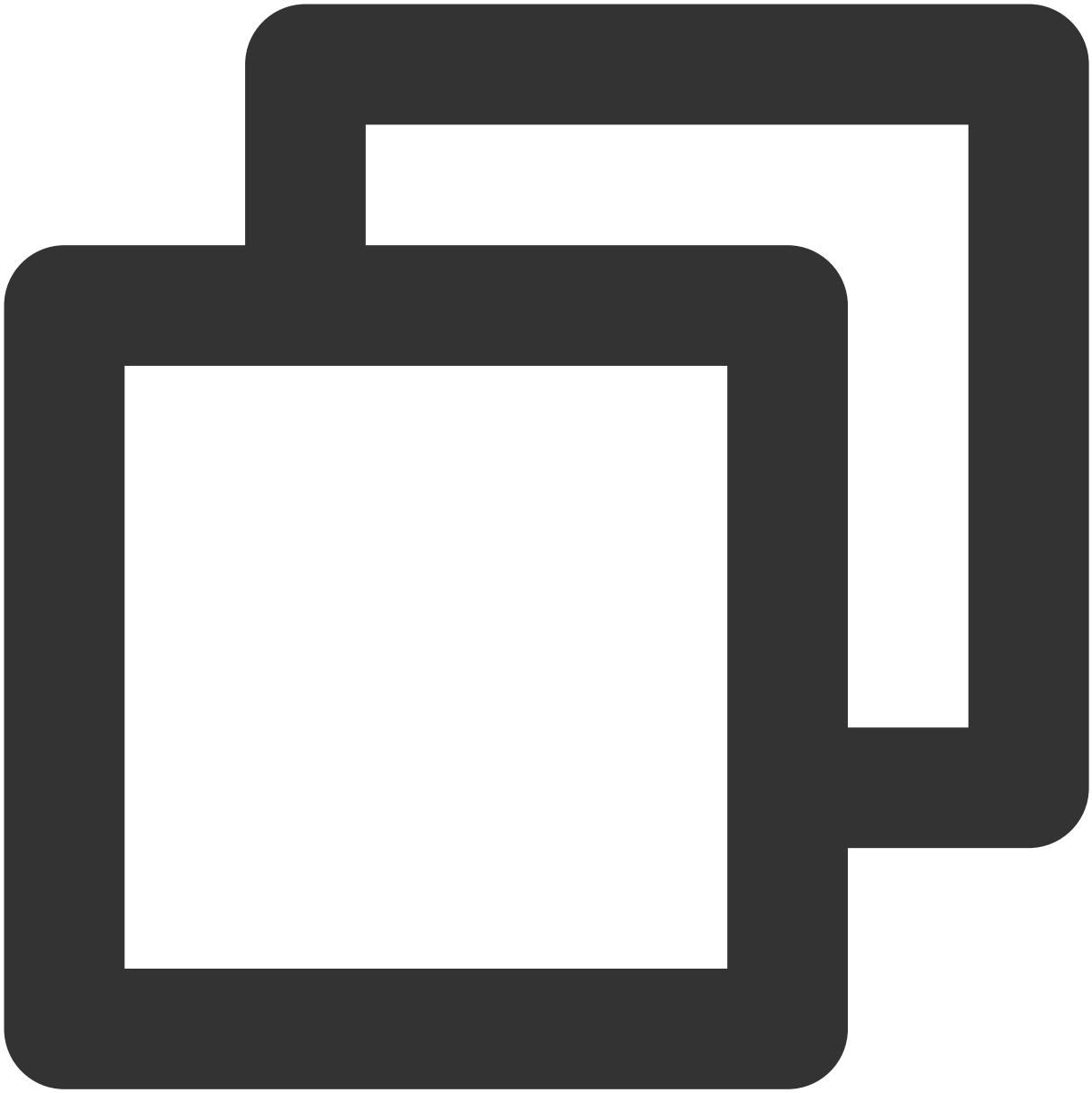
```
asr.cloud.tencent.com/asr/v2/1259228442?engine_model_type=16k_zh&expired=1592380492
```

2. 对签名原文使用 SecretKey 进行 HmacSha1 加密，之后再进行 base64 编码。例如对上一步的签名原文，SecretKey=kFpwoX5RYQ2SkqpeHgqmSzHK7h3A2fni，使用 HmacSha1 算法进行加密并做 base64 编码处理：



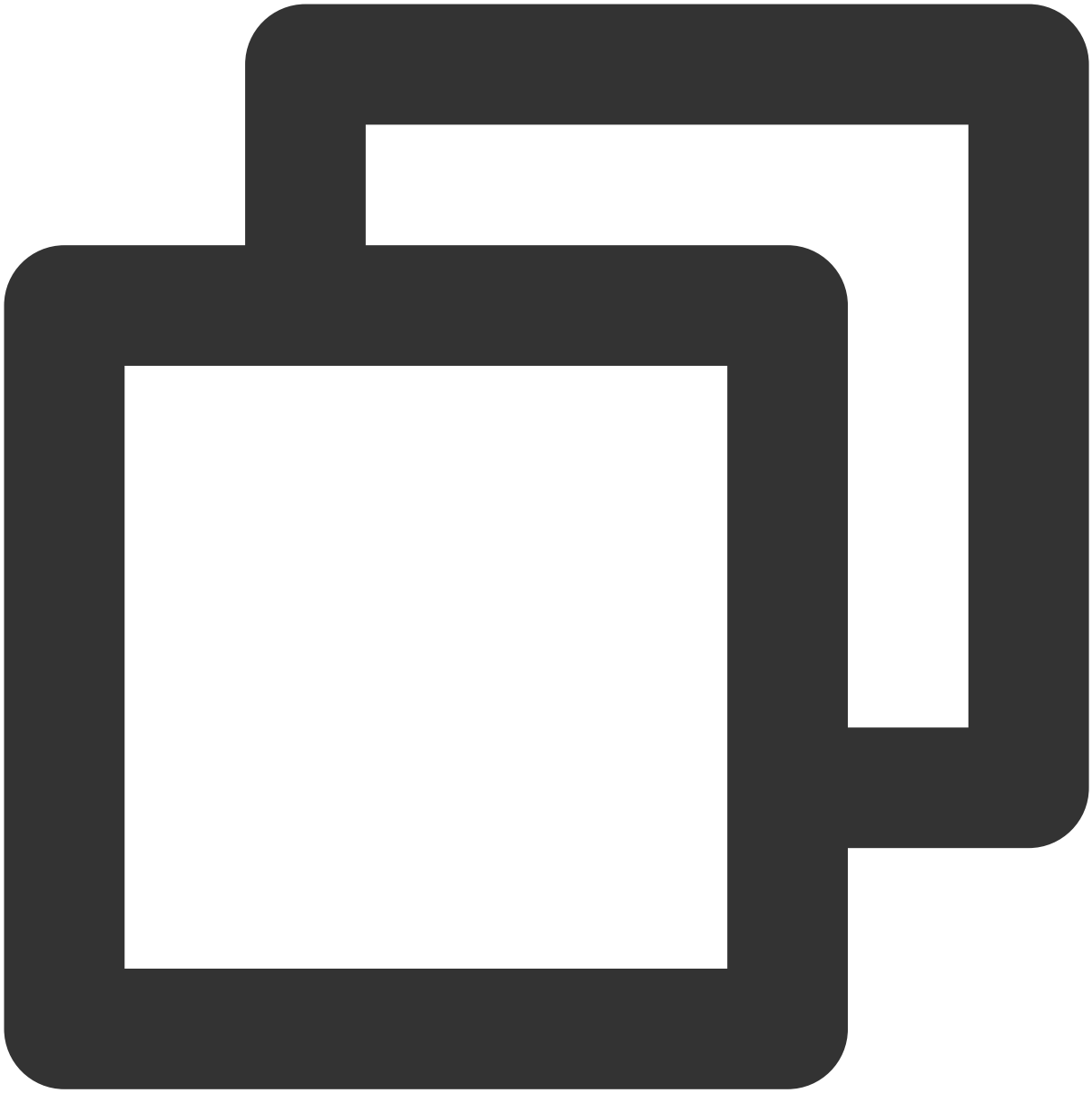
```
Base64Encode(HmacSha1("asr.cloud.tencent.com/asr/v2/1259228442?engine_model_type=16
```

得到 signature 签名值为：



```
HepdTRX6u155qIPKNKC+3U0j1N0=
```

3. 将 signature 值进行 **urlencode（必须进行 URL 编码，否则将导致鉴权失败偶现**）后拼接得到最终请求 URL 为：



```
wss://asr.cloud.tencent.com/asr/v2/1259228442?engine_model_type=16k_zh&expired=1592
```

Opus 音频流封装说明

压缩 FrameSize 固定640，即一次压缩640 short，否则解压会失败。传到服务端可以是多帧的拼接组合，每一帧需满足下面格式。

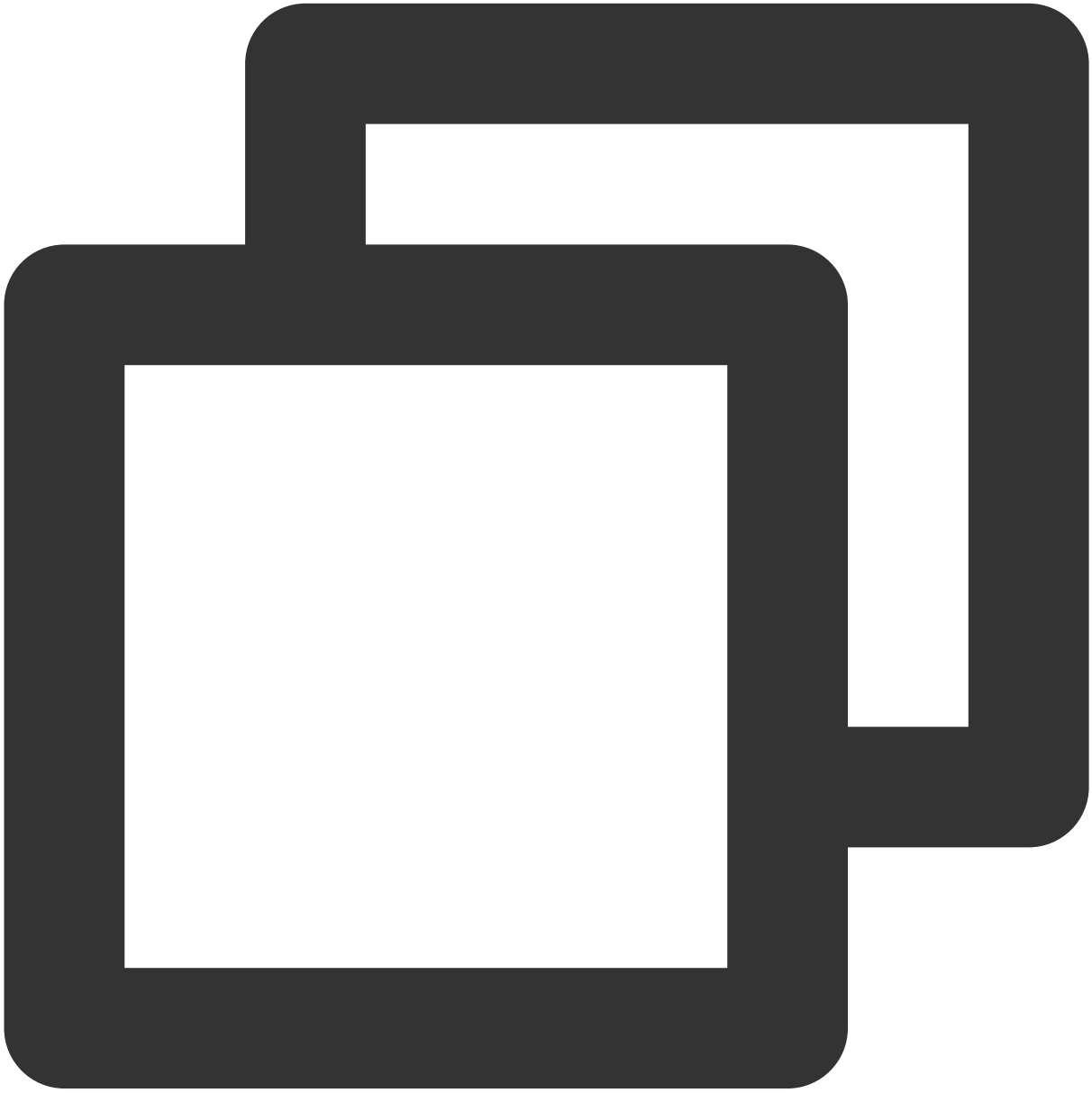
每一帧压缩数据封装如下：

OpusHead（4字节）	帧数据长度（2字节）	Opus 一帧压缩数据

opus	长度 len	对应 len 长的 opus decode data

请求响应

客户端发起连接请求后，后台建立连接进并进行签名校验，校验成功则返回 code 值为0的确认消息表示握手成功；如果校验失败，后台返回 code 为非0值的消息并断开连接。



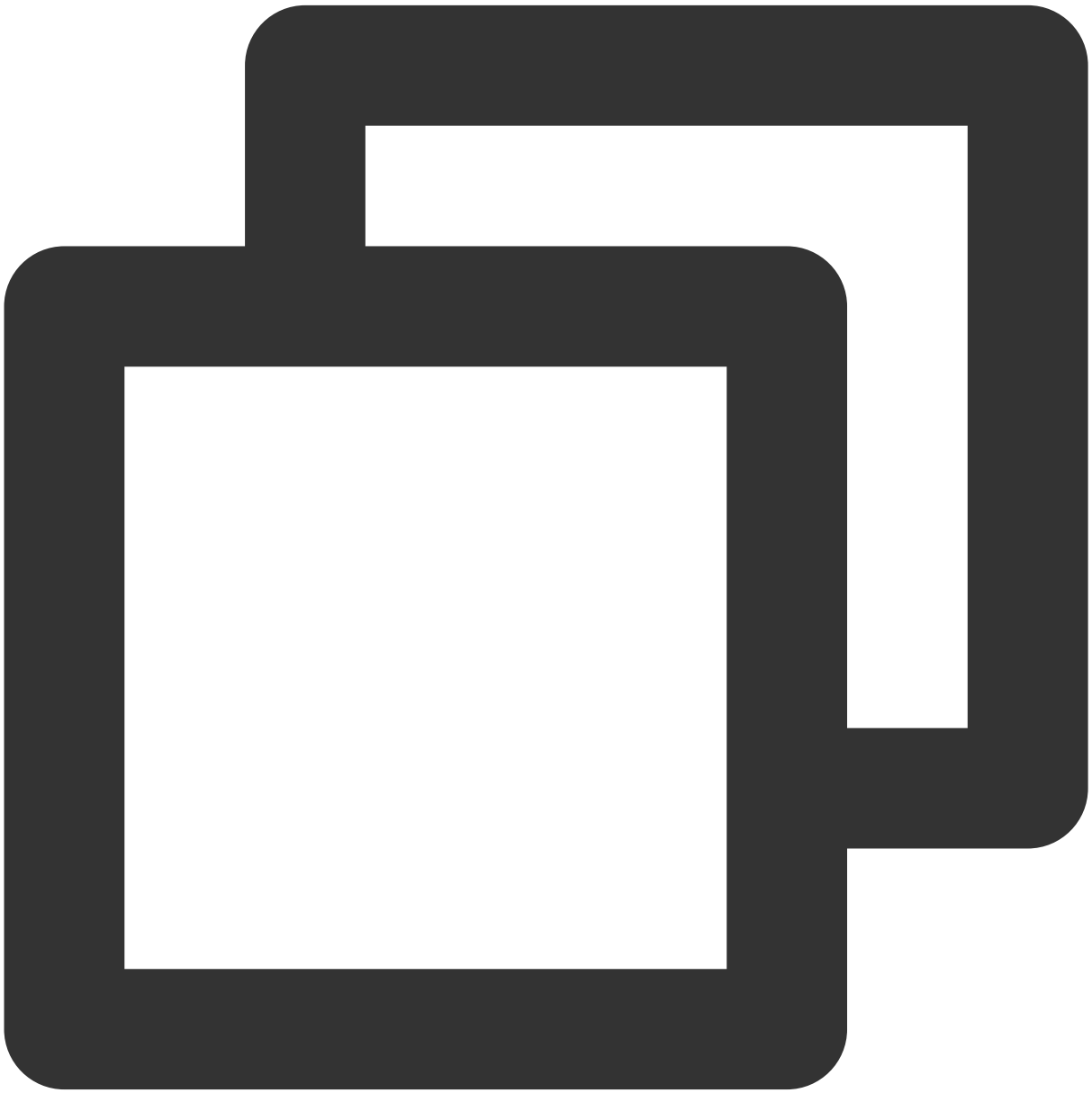
```
{"code":0,"message":"success","voice_id":"RnKu9FODFHK5FPpsrN"}
```

识别阶段

握手成功之后，进入识别阶段，客户端上传语音数据并接收识别结果消息。

上传数据

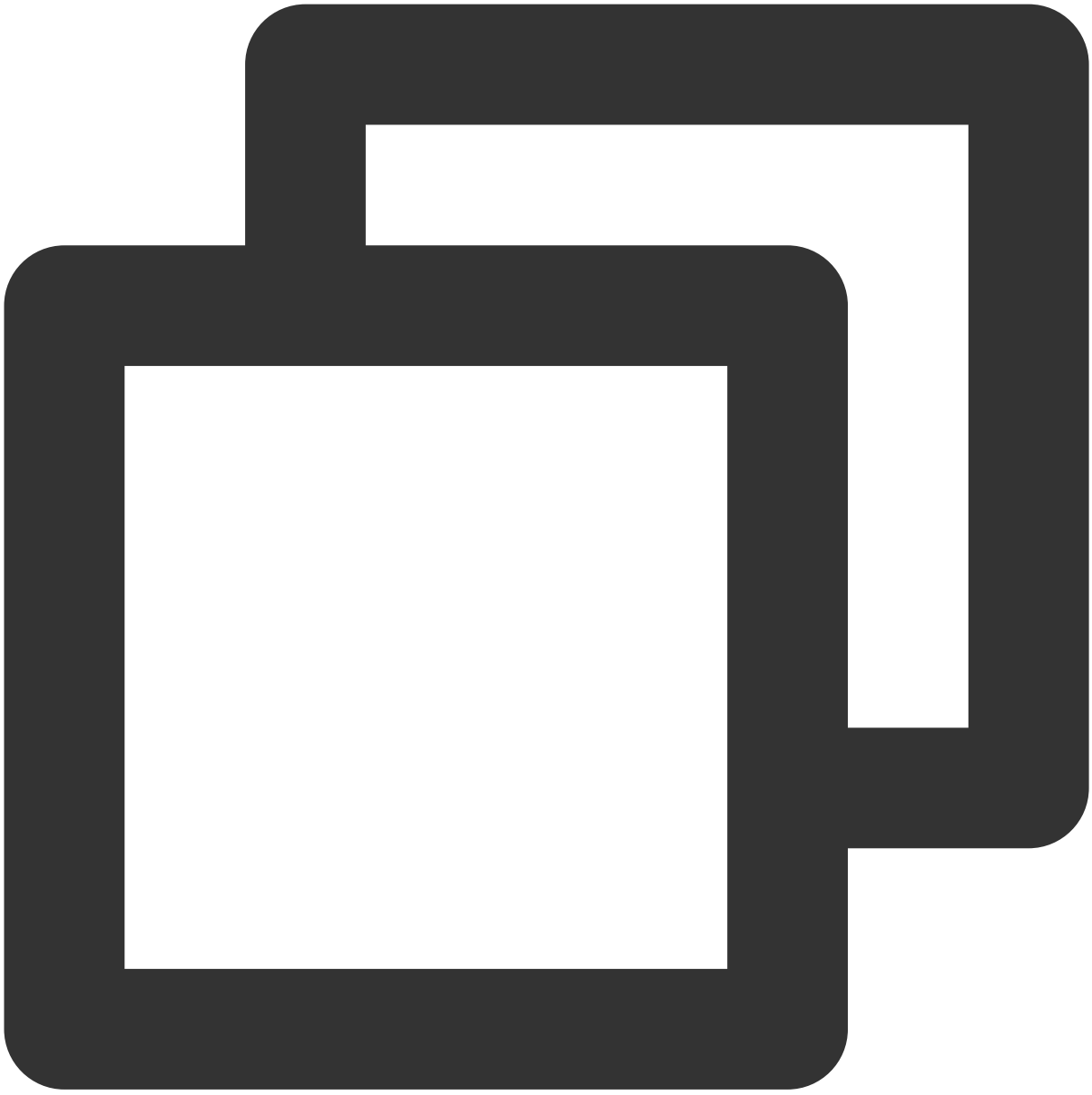
在识别过程中，客户端持续上传 binary message 到后台，内容为音频流二进制数据。建议每40ms 发送40ms 时长（即1:1实时率）的数据包，对应 pcm 大小为：8k 采样率640字节，16k 采样率1280字节。音频发送速率过快超过1:1实时率或者音频数据包之间发送间隔超过6秒，可能导致引擎出错，后台将返回错误并主动断开连接。
音频流上传完成之后，客户端需发送以下内容的 text message，通知后台结束识别。



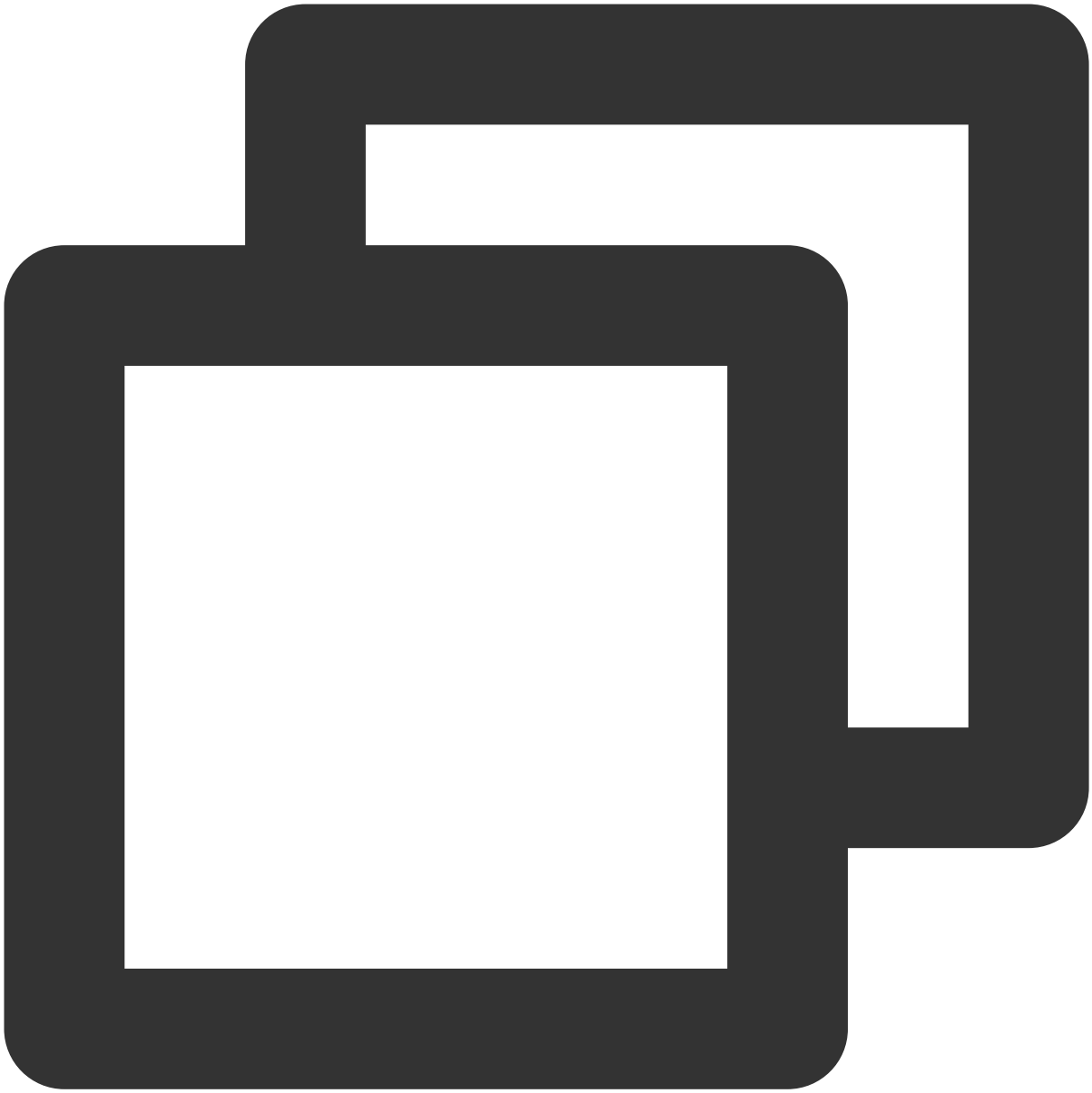
```
{"type": "end"}
```

接收消息

客户端上传数据的过程中，需要同步接收后台返回的实时识别结果，结果示例：

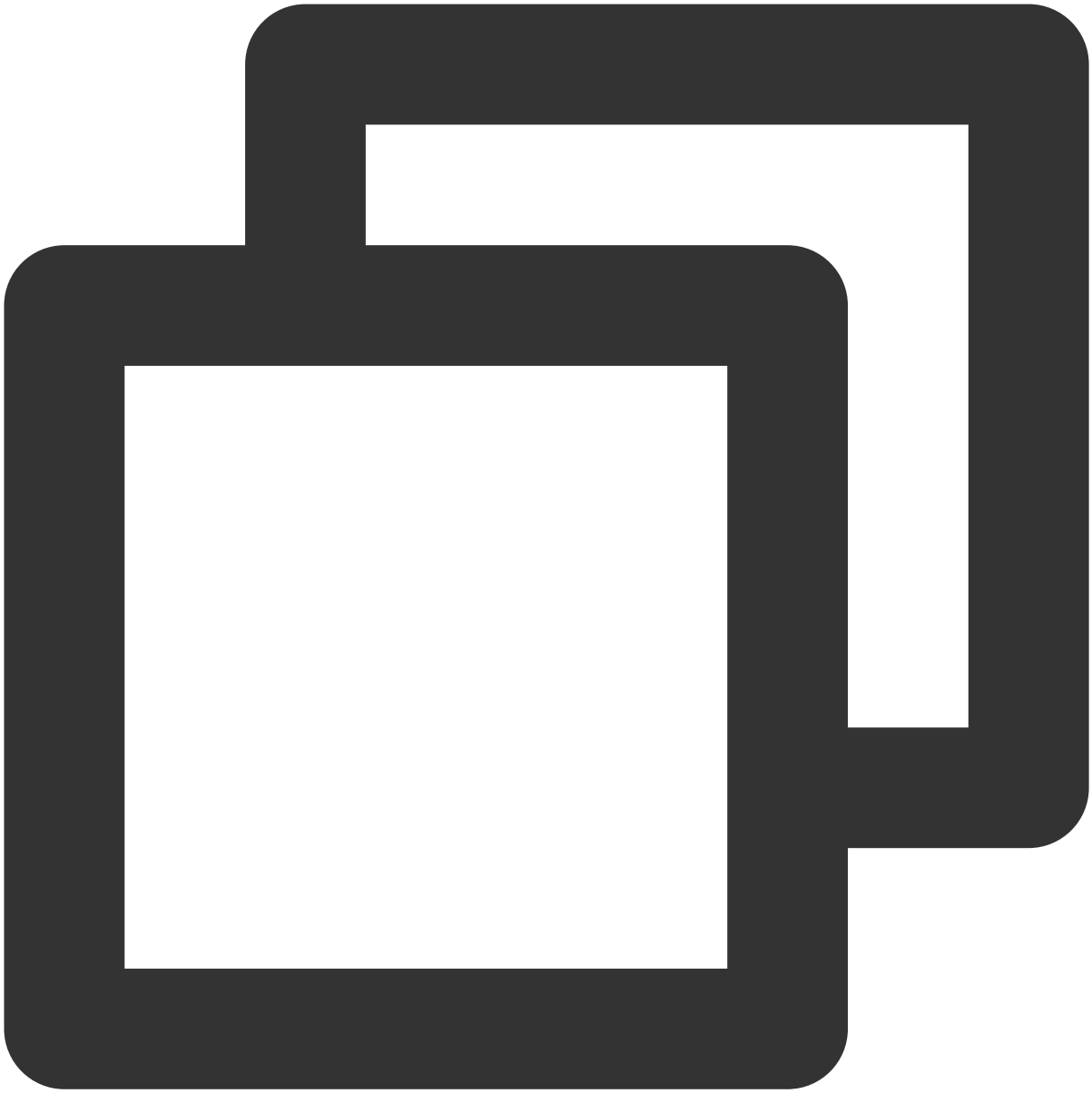


```
{"code":0,"message":"success","voice_id":"RnKu9FODFHK5FPpsrN","message_id":"RnKu9F
```



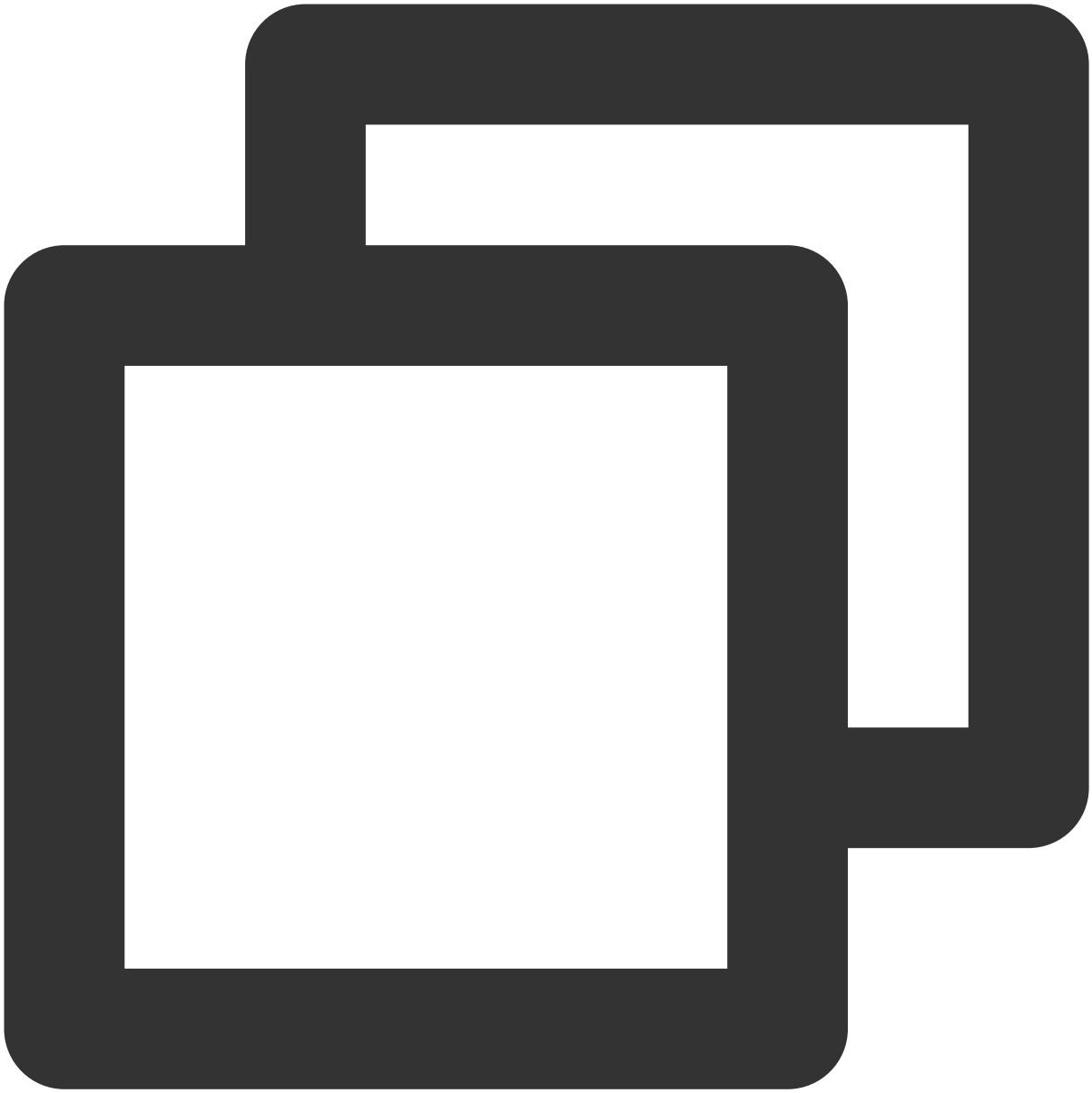
```
{"code":0,"message":"success","voice_id":"RnKu9FODFHK5FPpsrN","message_id":"RnKu9FO
```

后台识别完所有上传的语音数据之后，最终返回 final 值为1的消息并断开连接。



{"code":0,"message":"success","voice_id":"CzhjnqBkv8lk5pRUxhpX","message_id":"Czhjn

识别过程中如果出现错误，后台返回 code 为非0值的消息并断开连接。



{"code":4008,"message":"后台识别服务器音频分片等待超时","voice_id":"CzhjnqBkv8lk5pRUxhpX

开发者资源

SDK

Tencent Cloud Speech SDK for Go

Tencent Cloud Speech SDK for Java

Tencent Cloud Speech SDK for C++

Tencent Cloud Speech SDK for Python

Tencent Cloud Speech SDK for JS

SDK 调用示例

Golang 示例

Java 示例

C++ 示例

Python 示例

JS 示例

错误码

数值	说明
4001	参数不合法，具体详情参考 message
4002	鉴权失败
4003	AppID 服务未开通，请在控制台开通服务
4004	无可使用的免费额度
4005	账户欠费停止服务，请及时充值
4006	账号当前调用并发超限
4007	音频解码失败，请检查上传音频数据格式与调用参数一致
4008	客户端数据上传超时
4009	客户端连接断开
4010	客户端上传未知文本消息
5000	后台错误，请重试
5001	后台识别服务器识别失败，请重试
5002	后台识别服务器识别失败，请重试