

Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid)

Документация, содержащая информацию,
необходимую для эксплуатации экземпляра ПО

6.12.2023

Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid). Документация, содержащая информацию, необходимую для эксплуатации экземпляра ПО. Версия 1

Дата подготовки документа: 6.12.2023

Этот документ является составной частью технической документации Яндекса.

© 2008—2023 ООО «ЯНДЕКС». Все права защищены.

Предупреждение об исключительных правах и конфиденциальной информации

Исключительные права на все результаты интеллектуальной деятельности и приравненные к ним средства индивидуализации юридических лиц, товаров, работ, услуг и предприятий, которым предоставляется правовая охрана (интеллектуальную собственность), используемые при разработке, поддержке и эксплуатации службы Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid), включая, но не ограничиваясь, программы для ЭВМ, базы данных, изображения, тексты, другие произведения, а также изобретения, полезные модели, товарные знаки, знаки обслуживания, коммерческие обозначения и фирменные наименования, принадлежат ООО «ЯНДЕКС» либо его лицензиарам.

Использование результатов интеллектуальной деятельности и приравненных к ним средств индивидуализации в целях, не связанных с разработкой, поддержкой и эксплуатацией службы Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid), не допускается без получения предварительного согласия правообладателя. Настоящий документ содержит конфиденциальную информацию ООО «ЯНДЕКС». Использование конфиденциальной информации в целях, не связанных с разработкой, поддержкой и эксплуатацией службы Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid), а равно как и разглашение таковой, не допускается. При этом под разглашением понимается любое действие или бездействие, в результате которых конфиденциальная информация в любой возможной форме (устной, письменной, иной форме, в том числе с использованием технических средств) становится известной третьим лицам без согласия обладателя такой информации либо вопреки трудовому или гражданско-правовому договору.

Отношения ООО «ЯНДЕКС» с лицами, привлекаемыми для разработки, поддержки и эксплуатации службы Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid), регулируются законодательством Российской Федерации и заключаемыми в соответствии с ним трудовыми и/или гражданско-правовыми договорами (соглашениями). Нарушение требований об охране результатов интеллектуальной деятельности и приравненных к ним средств индивидуализации, а равно как и конфиденциальной информации, влечет за собой дисциплинарную, гражданско-правовую, административную или уголовную ответственность в соответствии с законодательством Российской Федерации.

Контактная информация

ООО «ЯНДЕКС»

<https://www.yandex.ru>

Тел.: +7 495 739 7000

Email: pr@yandex-team.ru

Главный офис: 119021, Россия, г. Москва, ул. Льва Толстого, д. 16

Содержание

- Эксплуатация.....4
 - Эксплуатация программы.....4
 - Начало работы.....4

Эксплуатация

Эксплуатация программы

Подробная информация о функционале сервиса представлена в [Справке](#).

Начало работы

В Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid) доступны приложения Яндекс Спичкит (Yandex SpeechKit) для синтеза и распознавания речи:

- Приложение для синтеза речи tts.
- Приложение для распознавания речи stt.
- Утилиты для нагрузочного тестирования.

Чтобы начать работу с сервисом, протестируйте развертывание приложений в [Docker-контейнерах](#) внутри вашей инфраструктуры.

Перед началом работы

1. Проверьте, что ваше оборудование соответствует [системным требованиям](#).
2. [Установите и настройте](#) необходимое окружение.
3. Установите на сервер [Docker Engine](#) и [Docker Compose](#).

Развертывание приложения синтеза речи

1. Создайте в отдельной директории файл `docker-compose.yml`, предназначенный только для тестирования развертывания приложения:

docker-compose.yml

```
version: '3'
services:
  tts:
    network_mode: host
    environment:
      - LICENSE_SERVICE_ENDPOINTS=0.0.0.0:9087
      - SERVICE_HOST=0.0.0.0
      - SERVICE_PORT=17009
      - UNISTAT_PORT=17003
      - GRPC_VERBOSITY=debug
      - LOGGING_LEVEL=INFO
      - TZ=UTC
      - NVIDIA_VISIBLE_DEVICES=0
      - NVIDIA_DRIVER_CAPABILITIES=compute,utility
    image: cr.yandex/идентификатор_пеестра/tts/v100/tts_server:'SK_Hybrid_version'
    deploy:
      resources:
        reservations:
          devices:
            - capabilities: ["gpu"]
              driver: nvidia
              device_ids: ["0"]

  license_server:
    network_mode: host
    environment:
      LICENSE_MODE: billing_agent
      LOGGING_LEVEL: INFO
      USE_TLS: "false"
      STATIC_API_KEY: 'API_KEY'
    image: cr.yandex/идентификатор_пеестра/license_server:'SK_Hybrid_version'

  envoy:
    network_mode: host
    image: cr.yandex/идентификатор_пеестра/envoy:'SK_Hybrid_version'
    environment:
      LOGGING_LEVEL: INFO
      ENVOY_UID: 0
    volumes:
      - ./log:/var/log/envoy
    depends_on:
```

```
- license_server
```

В файле укажите параметры, полученные при [настройке окружения](#):

- 'API_KEY' — идентификатор созданного API-ключа.
 - 'идентификатор_реестра' — идентификатор реестра Container Registry, в котором находятся подготовленные файлы приложения.
 - 'SK_Hybrid_version' — версия образа Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid).
2. Запустите приложение из директории с файлом `docker-compose.yaml`:

```
docker-compose up
```

Развертывание приложения распознавания речи

Примечание: Не рекомендуется разворачивать приложения синтеза и распознавания речи на одной ВМ, так как это создает большую нагрузку на систему.

1. Создайте в отдельной директории файл `docker-compose.yaml`, предназначенный только для тестирования развертывания приложения:

`docker-compose.yaml`

```
version: '3'
services:
  asr:
    network_mode: host
    privileged: true
    environment:
      - LICENSE_SERVICE_ENDPOINTS=0.0.0.0:8087
      - SERVICE_HOST=0.0.0.0
      - SERVICE_PORT=17002
      - UNISTAT_PORT=17003
      - GRPC_VERBOSEITY=debug
      - LOGGING_LEVEL=INFO
      - TZ=UTC
      - NVIDIA_VISIBLE_DEVICES=0
      - NVIDIA_DRIVER_CAPABILITIES=compute,utility
    image: cr.yandex/'идентификатор_реестра'/stt/v100/stt_server:'SK_Hybrid_version'
    deploy:
      resources:
        reservations:
          devices:
            - capabilities: ["gpu"]
              driver: nvidia
              device_ids: ["0"]

  license_server:
    network_mode: host
    environment:
      LICENSE_MODE: billing_agent
      LOGGING_LEVEL: INFO
      USE_TLS: "false"
      STATIC_API_KEY: 'API_KEY'
    image: cr.yandex/'идентификатор_реестра'/license_server:'SK_Hybrid_version'

  envoy:
    network_mode: host
    image: cr.yandex/'идентификатор_реестра'/envoy:'SK_Hybrid_version'
    environment:
      LOGGING_LEVEL: INFO
      ENVOY_UID: 0
    volumes:
      - ./log:/var/log/envoy
    depends_on:
      - license_server
```

В файле укажите параметры, полученные при [настройке окружения](#):

- 'API_KEY' — идентификатор созданного API-ключа.
- 'идентификатор_реестра' — идентификатор реестра Container Registry, в котором находятся подготовленные файлы приложения.
- 'SK_Hybrid_version' — версия образа Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid).

2. Запустите приложение для распознавания речи из директории с файлом `docker-compose.yml`:

```
docker-compose up
```

Нагрузочное тестирование

Чтобы оценить корректность и производительность тестовой инсталляции Яндекс Спичкит Гибрид (Yandex SpeechKit Hybrid), используйте контейнеры с утилитой нагрузочного тестирования для [синтеза](#) и [распознавания](#) речи.

1. Запустите нагрузочное тестирование синтеза речи на ВМ с приложением для синтеза речи:

```
docker run -it \
  --network=host \
  --env ENVOY_HOST="0.0.0.0" \
  --env ENVOY_TTS_PORT=9080 \
  --env USE_SSL=1 \
  --env RPS=1 \
  cr.yandex/'идентификатор_реестра'/tools/tts-tools:0.19
```

2. Запустите нагрузочное тестирование распознавания речи на ВМ с приложением для распознавания речи:

```
docker run -it \
  --network=host \
  --env SEND_MODE=RealTime \
  --env CONNECTIONS=30 \
  --env ENVOY_HOST="0.0.0.0" \
  --env ENVOY_PORT=8080 \
  --env USE_SSL=1 \
  --env ERRORS_THRESHOLD=0.1
  cr.yandex/'идентификатор_реестра'/tools/stt-tools:0.20
```

Если приложения работают успешно, на консоль выводятся сообщения с процентиями времени обработки запросов и другой информацией:

```
INFO: 2023-02-17 11:22:37.106 +0000 load_test.cpp:167 q=0.75: 250ms
INFO: 2023-02-17 11:22:37.106 +0000 load_test.cpp:167 q=0.9: 300ms
INFO: 2023-02-17 11:22:37.106 +0000 load_test.cpp:167 q=0.95: 300ms
INFO: 2023-02-17 11:22:37.106 +0000 load_test.cpp:167 q=0.99: 850ms
INFO: 2023-02-17 11:22:37.186 +0000 load_test.cpp:112 Total billing units: 260
INFO: 2023-02-17 11:22:37.186 +0000 load_test.cpp:218 SPS = 27.09172898
INFO: 2023-02-17 11:22:37.186 +0000 load_test.cpp:263 Errors statistics: 0% (0/50)
```