УТВЕРЖДЕН

ДССЛ.00142-01-ЛУ

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

«Зі Text To Speech»

Описание применения

ДССЛ.00142-01

Листов 16

2022

АННОТАЦИЯ

В данном документе приведено описание применения программного обеспечения (ПО) «3i Text To Speech» («3i TTS») и содержит описание программного интерфейса (API) для встраивания функции синтеза речи в онлайн сервисы, мобильные приложения и другие продукты.

В разделе 1 приводятся сведения о назначении и составе «3i TTS».

В разделе 2 указаны требования к программно-техническим средствам, необходимым для работы «3i TTS».

В разделе 3 указывается описание задач, решаемых «3i TTS», даются сведения об используемых технологиях.

В разделе 4 даются сведения об установке «3i TTS» и устранении основных ошибок в начале работы сервиса.

В разделе 5 приводится описание интерфейса программирования «3i TTS».

В разделе 6 даются сведения о входных и выходных данных «3i TTS».

СОДЕРЖАНИЕ

1. НАЗНАЧЕНИЕ ПРОГРАММЫ

Программное обеспечение «3i Text To Speech» предназначено для решения задачи синтеза речи, т.е. для преобразования данных из текстового вида в аудиопоток, содержащий его голосовой аналог. 3i TTS API позволяет произвести озвучку любого текста, используя технологию синтеза речи.

ПО состоит из одного модуля 3i-tts-service и реализован в виде серверного приложения, запускаемого в среде Docker.

2. УСЛОВИЯ ПРИМЕНЕНИЯ

2.1. Для функционирования ПО «3i TTS» необходима вычислительная система с доступными не менее:

- 2-х CPU (2 физических ядер) с частотой 2.4 GHz;
- одного GPU (графический ускоритель) NVIDIA с поддержкой CUDA и объемом ОЗУ не менее 8 ГБ;
- 8 ГБ ОЗУ;
- 30 ГБ дискового пространства.

2.2. Для функционирования ПО «3i TTS» на вычислительной системе должно быть установлено следующее общее программное обеспечение:

- 64-х разрядная операционная система на базе Linux (CentOS не ниже 7.X, Debian не ниже 10.x , Astra Linux не ниже 1.7);
- docker, nvidia-docker, docker-compose;
- драйвер cuda версии не ниже10.2;
- python3 (для работы тестового приложения).

3. ОПИСАНИЕ ЗАДАЧИ

Основная задача, решаемая при помощи программного обеспечения (ПО) «3i Text To Speech» - генерация аудиопотока для входной текстовой строки максимально близкой по звучанию к человеческой речи. «3i TTS» относится к системам параметрического синтеза. Достоинствами параметрического синтеза являются естественность звучания (особенно при нейросетевых подходах), большее разнообразие в интонациях и использование меньшего объема обучающих данных. В ПО «3i TTS» синтез речи осуществляется посредством двух основных этапов – (1) преобразования текста в спектрограмму и (2) преобразования спектрограммы в аудиопоток. На этапе (1) используются такие подходы, как рекуррентные нейронные сети (RNN – Recurrent neural network) и seq2seq. На этапе (2) используется вокодер на основе архитектуры WaveNet

Синтез может применяться для задач:

- озвучки аудиокниг, видеороликов, объявлений, навигационных систем, игрушек и т.д.
- создания роботов и виртуальных ассистентов;
- информационно-справочных систем для помощи слепым и немым;
- автоматического оповещения по каналам связи и т.д.

Качество синтеза

Основной метрикой качества синтеза является MOS (mean opinion score), усредненная оценка естественности речи, определенная человеком (ассессором) для синтезированных аудиоданных по шкале от 1 до 5. Единица относит аудио к неправдоподобному звучанию, а пятерка — речь, неотличимая от человеческой. Современные системы синтеза речи имеют оценку выше 4.

4. ВЫЗОВ И ЗАГРУЗКА ПРОГРАММЫ

4.1 УСТАНОВКА ПО «3I TTS»:

Процесс установки ПО «3i TTS» состоит из следующих шагов:

1. Убедиться в наличии необходимой версии драйверов CUDA:

```
nvcc --version
```

2. Установить на сервер docker с сайта docker.com. Для работы на GPU также установить nvidia-docker.
3. Авторизоваться в репозитории от 3itech

```
sudo docker login docker-registry.3itech.ru

username: products
password: *******
```

4. Скачать образ

```
sudo docker pull docker-registry.3itech.ru/production/3i-tts-service:v1.0.0
```

5. Создать docker-compose.yaml

```
version: '2.3'

 services:
   3i-tts-service:
     image:  docker-registry.3itech.ru/production/3i-tts-service:v1.0.0
     container_name: 3i-tts-service
     restart: always
     runtime: nvidia
     environment:
       - NVIDIA_VISIBLE_DEVICES=0
     volumes:
         - "/opt/docker-mounts/text-to-speech/Logs:/text-to-speech/Logs"
     ports:
       - 60001:60001
```

Красным цветом отмечены параметры, которые можно изменить, а именно:

- номер видеокарты, на котором будет запущен сервис,
- путь к каталогу с лог-файлами;
- порт, на который будут отправляться запросы.

6. Запустить сервис

```
docker-compose up -d
```

7. Проверить работоспособность:

a) Установить зависимость:

```
python3 -m pip install -U grpcio_tools
```

b) Скачать тестовое приложение

```
https://cloud.3i-tp.ru/d/e32792f98a0840408566/
pass: *******
```

c) Запустить тестовое приложение

```
cd client_data
python3 tts_client.py
```

d) Удостовериться в наличии файла «test.wav» в каталоге с тестовым приложением.

4.2 ОСНОВНЫЕ СООБЩЕНИЯ ОПЕРАТОРУ:

Логи работы ПО «3i TTS» можно получить либо из монтированного из контейнера файла «user.log», либо с помощью команды:

```
docker logs <имя контейнера>
```

Успешный старт моделей:

```
INFO     | normalizer:<module>:23 - Normalizer is ready
INFO     | modules.tts.synthesizer:_load_text_handler:282 - Loading text handler
INFO     | modules.tts.synthesizer:module_from_config:241 - Loading tacotron2 module
INFO     | modules.tts.synthesizer:module_from_config:241 - Loading waveglow module
INFO     | modules.tts.synthesizer:load_user_dict:189 - User dictionary has been loaded
INFO     | modules.tts.synthesizer:__init__:65 - Synthesizer female is ready
```

Логгирование запроса:

```
INFO    | modules.tts.synthesizer:synthesize:69 - Этот пример запроса с текстом, который будет синтезирован
INFO    |    modules.tts.synthesizer:save:152    -    Audio    was    saved    as    /text-to-speech/data/waves/female_3b0a87d9497746_2022-08-30_11-50.wav
```

4.3 ОСНОВНЫЕ ОШИБКИ

При запуске образа возможны следующие ошибки:

1) Отсутствие поддержки nvidia-runtime со стороны некоторых версий docker-compose.

Исправляется установкой нужной версией docker-compose:

```
pip3 install docker-compose==1.23.0
```

2) RuntimeError: CUDA error: out of memory

Исправляется выделением достаточного количества памяти на GPU в соответствии с указанными аппаратно-программными требованиями

5. ОПИСАНИЕ ФУНКЦИЙ API

API сервиса реализовано при помощи протокола gRPC. Proto-файл с используемыми классами и методом запроса:

```
syntax = "proto3";

package vox.tts;

message SynthesizeRequest {
  string model = 1;
  string text = 2;
  float speed = 3;
  float tone = 4;
}

message SynthesizeResponse {
  bytes audio_content = 1;
}

service TTS {
    rpc synthesize(SynthesizeRequest) returns (SynthesizeResponse);
}
```

Список функций и структур API «3i TTS» представлен в таблице 5.1.

Таблица 5.1 Список функций и структур API 3i TTS

Наименование	**Описание функции**
synthesize	Функция синтеза речи по входной текстовой строке: – на вход принимает структуру *SynthesizeRequest;* – возвращает структуру *SynthesizeResponse.*
SynthesizeRequest	Структура запроса, которая создает задачу на синтез. Параметры: – *model* – наименование модели, выбирается из доступных (напр. “male” – “мужская)”; – *text* – текст в кодировке UTF-8,

Наименование	Описание функции
	предназначенный для синтеза; для постановки ударения перед требуемой гласной необходимо поставить символ ‘+’, напр. “королевский з+амок” или “дверной зам+ок”; – *speed* – коэффициент скорости воспроизведения, число в диапазоне [0.5, 2.0]; – *tone* – коэффициент высоты звука, число в диапазоне [0.75, 1.5].
SynthesizeResponse	*audio_content* – буфер с аудиопотоком в формате WAV и следующими характеристиками: – частота дискретизации 22050 Гц; – кодек pcm_s16le; – количество каналов 1.

6. ВХОДНЫЕ И ВЫХОДНЫЕ ДАННЫЕ

Входными данными для «3i TTS» является строка в кодировке UTF-8, которая может содержать управляющие символы.

Выходными данными является буфер, содержащий аудиоданные с характеристиками:

- частота дискретизации 22050 Гц;
- кодек pcm_s16le;
- количество каналов 1.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ

В настоящем документе приняты следующие условные обозначения:

ПО	Программное обеспечение
TTS	сокр. англ.Text To Speech преобразование текста в речь
3i TTS	ПО «3i Text To Speech»
API	сокр. англ. Application Programming Interface, интерфейс программирования приложений – набор готовых классов, процедур, функций, структур и констант, предоставляемых приложением для использования во внешних программных продуктах.
gRPC	сокр. англ. Remote Procedure Call – высокопроизводительный протокол удаленного вызова процедур.
CPU	сокр. англ. Central Processing Unit – электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (код программ).
RNN	сокр. англ. Recurrent neural network, искусственная нейронная сеть, где элементы образуют направленную последовательность.
seq2seq	сокр. англ. sequence to sequence (из последовательности в последовательность), семейство подходов машинного обучения, используемого для задач обработки естественного языка.
WaveNet	глубокая нейронная сеть для генерации необработанного звука.

ОЗУ	сокр. Оперативное запоминающее устройство – реализация функции оперативной памяти.
GPU	сокр. англ. graphics processing unit, графический ускоритель.

Лист регистрации изменений									
Изм.	Номера листов (страниц)				Всего листов (страниц) в документе	№ доку мента	Входящий № сопроводи тельного документа и дата	Под пись	Дата
	Изме-нен-ных	заменен ных	новых	аннули рован ных					
