

Что такое речь

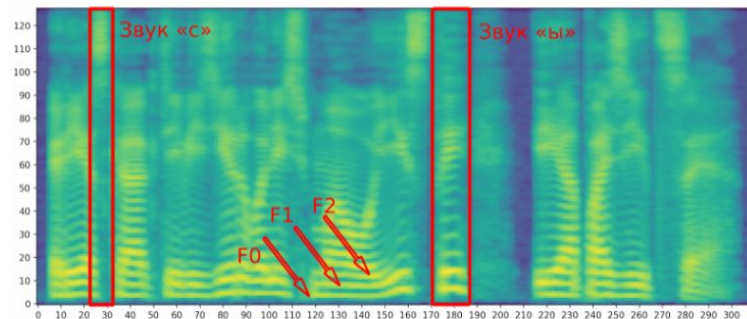
<https://habr.com/ru/company/sberdevices/blog/548812/>

В задачах speech processing лучше пользоваться time-frequency-представлением звука с помощью спектрограмм (short time Fourier transform, STFT). Математически это временная последовательность модулей преобразования Фурье от коротких (10-20 мс) отрезков звука, внутри которых сигнал можно считать стационарным, то есть его спектральные характеристики почти не меняются за это время. Причины того, почему такой подход работает, тоже можно найти в биологии речевого тракта.



Речевой тракт человека

Человек разговаривает с помощью голосовых связок и других органов речи. Воздух выдыхается из легких, колеблет мембраны голосовых связок, получается периодический сигнал. Затем он резонирует, проходит через несколько фильтров (горло, нёбо, язык, зубы, губы), обрастает дополнительными гармониками (модулируется) и выходит изо рта в таком виде, в каком мы его слышим. Голосовые связки — это не главный орган речи человека. Например, они никак не участвуют при произнесении глухих согласных — с, п, к, ... На спектрограмме они выглядят как высокочастотные равномерно раскрашенные области, а вокализованные звуки (все гласные и звонкие согласные) — как несколько ярких полос, с наибольшей амплитудой в низкочастотной области (в нижней части спектрограммы). Самая первая (нижняя) полоса называется fundamental frequency (частота основного тона, F0) — это и есть частота колебаний голосовых связок. Следующие гармоники (полосы F1, F2, ...) могут иметь большую амплитуду, но кратны F0.



Мел-спектрограмма 4-секундного аудио.

На мел-спектрограммах каждый столбец на ней представляет собой rFFT от короткого фрагмента аудио. По оси X отложено время, по Y — номер мел-фильтра. Мел-шкала — это такой способ снизить разрешение спектрограмм по частоте с 2000 до 128 (или даже 80) без особенной потери информации. Он основан на психоакустике: восприятие человеком высоты и громкости звука логарифмическое. То есть нам кажется, что звук стал выше на какую-то величину, когда в действительности высота звука выросла в какое-то количество раз. Более подробно про процессинг мел-спектрограмм можно почитать [тут](#).

<https://habr.com/ru/company/tinkoff/blog/474782/>

Параметрический синтез речи

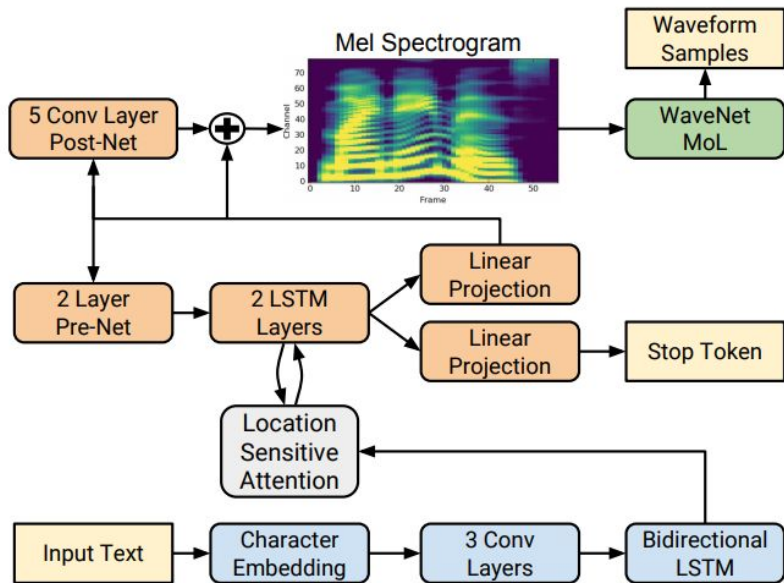


Схема Tacotron 2.

Статья на Хабре от silero.ai

<https://habr.com/ru/post/563484/>

Если коротко:

- Мы сделали наш вокодер в 4 раза быстрее;
- Мы сделали пакетирование моделей более удобным;
- Мы сделали мультиспикерную / мультязычную модель и "заставили" спикеров говорить на "чужих" языках;
- Мы добавили в наши русские модели возможность автопростановки ударений и буквы ё с некоторыми ограничениями;
- Теперь мы можем сделать голос с нормальным качеством на 15 минутах — 1 часе (с теплого старта в принципе заводилось даже на 3-7 минутах) или на 5 часах аудио (с холодного старта). Но тут все очень сильно зависит от качества самого аудио и ряда деталей;
- Мы привлекли комьюнити к работе, и нам помогли сделать удобный интерфейс для записи. Мы начали работу над голосами на языках народностей СНГ (украинский, татарский, башкирский, узбекский, таджикский). **Если вы хотите увидеть свой язык в числе спикеров — пишите нам;**
- Мы продолжаем собирать обратную связь по применимости нашей системы для экранных интерфейсов чтения, и пока кажется, что нужно где-то еще всё ускорить в 5-10 раз, чтобы наши модели закрывали и этот кейс;

Запись данных



Канкаев Эрдни сделал **2660** записей общей длительностью около **5,5 часов**

Пример:

Эмгн өвгн хойр тагчг, дэхэд эдн ю келдгчнь гиһэд, ө уга сууцхав.



Запись данных



Speaker: delghir
Project: kalmyk_fairytales

⚙️ Allow recording

Көвүд, күүкд ирэд: Аав, ээж юунд уульдз йовхмт?.



▶ Start recording

■ Stop recording

▶ Play

⬇ Download

🗑 Delete

00:00

Характеристики Микрофон Fifine K669B



Технические характеристики

Чувствительность	-34 дБ
Минимальная частота	20 Гц
Максимальная частота	20000 Гц
Максимальный уровень звукового давления	130 дБ
Соотношение сигнал/шум	78 дБ

Анонсирование синтеза калмыцкого языка от silero.ai

Новости нашего синтеза

Публичные голоса народов СНГ

Вместе с комьюнити мы сделали и опубликовали полностью [уникальные модели](#) языков народов СНГ:

- Башкирский (`aigul_v2`);
- Калмыцкий (`erdni_v2`);
- Татарский (`dilyara_v2`);
- Узбекский (`dilnavoz_v2`);

Мы также попробовали сделать украинский голос на публичных данных (из аудиокниг), но там получилось весьма посредственное качество (все остальные голоса люди записали с нуля).

Некоторые модели звучат почти идеально, некоторые похуже. Обычно это связано со стабильностью дикции. Но поскольку дикторы участвовали в этом на общественных началах, сложно было приставлять к ним "войс-коучей" и вообще стоять над душой.


На каждый голос мы использовали от 1 до 6 часов записей. Это модели без автоматической протановки ударения, они чуть быстрее как и все V2 модели.

К сожалению пока публичного украинского языка не будет, но просто в качестве дразнилки, вот пример того как это может звучать (автор голоса не разрешил нам публиковать модель) на голосе профессионального диктора:

Models and Speakers

All of the provided models are listed in the [models.yml](#) file. Any meta-data and newer versions will be added there.

Currently we provide the following speakers:

Speaker	Auto-stress	Language	SR	Colab
<code>aidar_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>baya_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>irina_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>kseniya_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>natasha_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>ruslan_v2</code>	yes	ru (Russian)	8000 , 16000	 Open in Colab
<code>lj_v2</code>	no	en (English)	8000 , 16000	 Open in Colab
<code>thorsten_v2</code>	no	de (German)	8000 , 16000	 Open in Colab
<code>tux_v2</code>	no	es (Spanish)	8000 , 16000	 Open in Colab
<code>gilles_v2</code>	no	fr (French)	8000 , 16000	 Open in Colab
<code>multi_v2</code>	no	ru , en , de , es , fr , tt	8000 , 16000	 Open in Colab
<code>aigul_v2</code>	no	ba (Bashkir)	8000 , 16000	 Open in Colab
<code>erdni_v2</code>	no	xal (Kalmyk)	8000 , 16000	 Open in Colab
<code>dilyara_v2</code>	no	tt (Tatar)	8000 , 16000	 Open in Colab
<code>dilnavoz_v2</code>	no	uz (Uzbek)	8000 , 16000	 Open in Colab
<code>mykyta_v2</code>	no	ua (Ukrainian)	8000 , 24000 , 48000	 Open in Colab

(!!!) In `multi_v2` all speakers can speak all of languages (with various levels of fidelity).

Пример запуска

```
▶ #@title model load
import torch

language = 'xal'
speaker = 'erdni_v2'
sample_rate = 16000
device = torch.device('cpu')
model, example_text = torch.hub.load(repo_or_dir='snakers4/silero-models',
                                     model='silero_tts',
                                     language=language,
                                     speaker=speaker)

model.to(device) # gpu or cpu
```

Using cache found in /root/.cache/torch/hub/snakers4_silero-models_master

```
[3] #@title default example
audio = model.apply_tts(texts=[example_text],
                        sample_rate=sample_rate)

print(example_text)
display(Audio(audio[0], rate=sample_rate))
```

horvn, dörvn күн ирэд, hazanь чиңгнв. Байн Цецн хаана horvn көвүн күүнджәнә.

▶ 0:00 / 0:08 🔊 ⋮

```
▶ #@title example №1
example_text_1 = 'Эмгн өвгн хойр тагчг, дэкэд эдн ю келдгчнь гинэд, э уга сууцхав'
audio_1 = model.apply_tts(texts=[example_text_1],
                          sample_rate=sample_rate)

print(example_text_1)
display(Audio(audio_1[0], rate=sample_rate))
```

▶ Эмгн өвгн хойр тагчг, дэкэд эдн ю келдгчнь гинэд, э уга сууцхав.

▶ 0:00 / 0:06 🔊 ⋮

https://github.com/AndTseren/Kalmyk_TTS/blob/main/example_kalm_ipynb%22.ipynb

https://github.com/AndTseren/Kalmyk_TTS/tree/main/results/erdni_v2/folklore/fairy_tales