Development of MVP Service for Speech Synthesis in English

Kalinina Daria 204

27 July 2022

Abstract

The purpose of the work was to study the technologies used in the field of speech synthesis and the creation of MVP text-to-speech service. The service was implemented as a telegram bot with the functions of translating a text query into an audio file and displaying analytics. Several microservices are connected to the telegram bot, which are responsible for each request and the stage of text processing.

Contents

Chapter 1

Theoretical part

1.1 Introduction

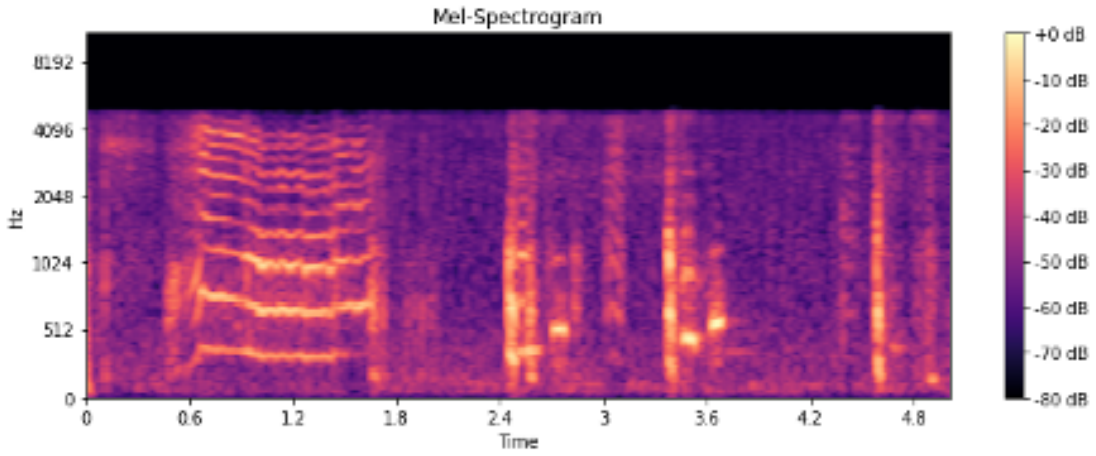
The scope of our project is speech synthesis or text-to-speech technology. This is a computer simulation of human speech from a text representation using machine learning methods. Speech synthesis, usually, is used by for phone robots, voice assistant and content voiceover.

Development of the technology consists of two stages: NLP-frontend and audioprocessing. The first stage is the NLP-frontend. At this step, the developer prepares the text for further processing. This preparation includes the normalization of the text and the translation of the grapheme into phonemes. Text normalization is bringing all the characters of the text from different semiotic classes into a single form: replacing numbers in digital notation with words in correct form, replacing abbreviations with the full form and special symbols with conversational analogs.

The next stage of text preprocessing is the translation of a grapheme into a phoneme and the insertion of accents. At this stage, it is necessary to transcribe the sounds correctly. In this case, it may be difficult to transcribe words that do not contain slang words in dictionaries. The translation of such graphemes is carried out using neural networks. Also, another ambiguity arises in the affixing of accents and sounds in homographs, where context is needed.

Examples: dessert, prEsent and presEnt.

After preprocessing, the data is converted into an audio track. This stage includes creating a MEL spectrogram using a tacotron, creating audio based on the spectrogram using a hi-fi gan and improving audio.



Almost every major company in the field of IT technology has its own text-to-speech service. They are most often used in translation programs, as voice assistants in search engines and companies providing services, such as banks. Here is a comparative table of some of these services.

	Google cloud text-to-speech	Amazon's Polly
Real-time audio transmission	+	+
SSML	+	+
Custom Voice	+	-

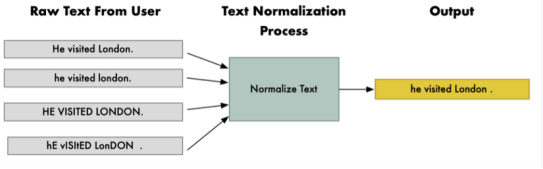
1.2 Stages of processing

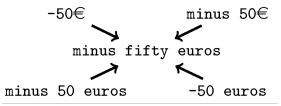
1.2.1 Normalization

The main task of normalization is to bring the text to a standard unified form. This is necessary for further processing steps so that the acoustic model can unambiguously interpret and transform the text. Thus, we bring our text closer to the form in which it is used in oral speech. Usually, abbreviations, special symbols and numbers are subject to normalization.

Examples of text that need to be normalized:

- 1917 → nineteen seventeen
- kg → kilogram
- 10:00 → ten o'clock . . .





Finite State Transducer is a graph that has an initial and final state. The transducer takes input data, in our case, raw text, and produces output – normalized text. Each state of the transducer is connected by paths to a pair of input and output and, depending on the input, moves between states. Thus, WFST has a weighted sum on each edge, estimating the probability of issuing one or another result.

1.2.2 Acoustic

The next step after NLP processing is the conversion of graphemes into mel-spectrograms. Such transformations from raw input text in our service are carried out by FastPitch, a fully feedforward Transformer model. The model consists of a bidirectional transducer, a pitch predictor, and a duration predictor. After passing through the first *N* blocks of the converter, encoding, the signal is supplemented with information about the pitch and discretely upsampled. It then goes through another set of *N* transformer blocks to smooth the upsampled signal and build a chalk spectrogram.

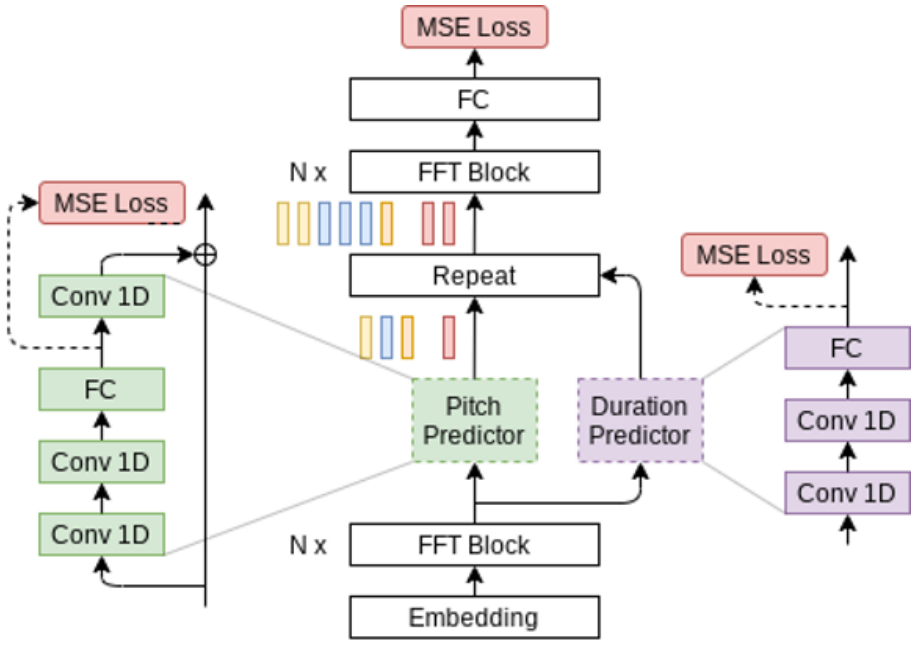


Figure 1.1: Acoustic model

The model is parallel, that allows you to quickly synthesize high-precision small-scale spectrograms with a high degree of control over speech parameters. The model provides the ability to adjust the tempo, expressiveness, pitch and intonation, which allows you to make the synthesized speech more natural.

Chapter 2

Bonus part

This bonus part is about everything and nothing concrete))

2.1 Genshin Impact

This is my top 5 favourite characters is Genshin impact.

1. Xiao
2. Tartaglia
3. Zhongli
4. Kaeya
5. Kazuha

This is list of items you shoud farm for you character to make him stringer.

1. Weapon
2. Artifacts
   (a) Flower
   (b) Feather
   (c) Watch
   (d) Cup
   (e) Crown
      i. Crit rate
      ii. Crit Damage
3. Books for talents
4. Items from bosses

For dipper information you can check this Xiao guide

Tier list of **main DD** [1] in Genshin.

S	Hu Tao	Ganyu	Ayaka
A	Xiao	Diluc	Yanfei
B	Kequing	Kaeya	Noelle

2.2 Some math

$$sin(\alpha + \beta) = sin\alpha cos\beta + cos\alpha sin\beta \tag{2.1}$$

$$\frac{n!}{k!(n-k)!} = \binom{n}{k} \tag{2.2}$$

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4}}}} \tag{2.3}$$

$$p_n(x) = p_{n-1}(x) + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \tag{2.4}$$

"Mathematics is the key and the door to all sciences." © Galileo Galilei

[1] damage dealer

Bibliography

[1] Eric Engelhart, Mahsa Elyasi, Gaurav Bharaj "Grapheme-to-Phoneme Transformer Model for Transfer Learning Dialects"

[2] Jungil Kong, Jaehyeon Kim, Jaekyoung Bae "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis"

[3] Evelina Bakhturina, Yang Zhang, Boris Ginsburg "Shallow Fusion of Weighted Finite-State Transducer and Language Model for Text Normalization"