## Inroduction

Investigations on text-to-speech (TTS) task are not stopping for decades and the problem is still treated being unsolved. By the way, eventually, with the rise of deep learning and, particularly, neural networks, research community has got closer to the natural high-fidelity human voice modelling. Nonetheless, a lot of issues remain unclosed. With the growing power of neural networks most of the solutions got unacceptably huge and overparameterized without having any possibility to be deployed on-device, which is crucial in building production voice assistants nowadays. Moreover, it is worth to notice that modelling such a complex time sequence with autoregression nature as audio waveform is not a trivial task at all. Thus, to build a personal voice assitant one should take care of many things: balanced ratio of parameter space size and model complexity, fast inference speed, etc.

In this paragraph we are going to recall the architecture of Tacotron 2 [1] – state-of-the-art model for TTS purposes, which we are going to compress. Being a sequence-to-sequence type of neural network it has encoder-decoder structure with attention mechanism in the middle. Encoder module consists of 512-dimensional embedding layer, 3 convolutions with kernel size of (5x1) and 512 channels and bidirectional LSTM, finally. Tensor of size (BxTx512) as Encoder output is filled to the attention module input, which aligns text with generating frame-by-frame mel-spectrogram. The crucial role in generation play 2 decoding recurrent units after attention with the final linear projection.

Comparing to the alternatives such as Transformer-TTS [2], Tacotron 2 has the most reasonable synthesis quality w.r.t. the volume of parameter space (28 milliom parameters of Tacotron 2 against 40 millions parameters of Transformer-TTS).