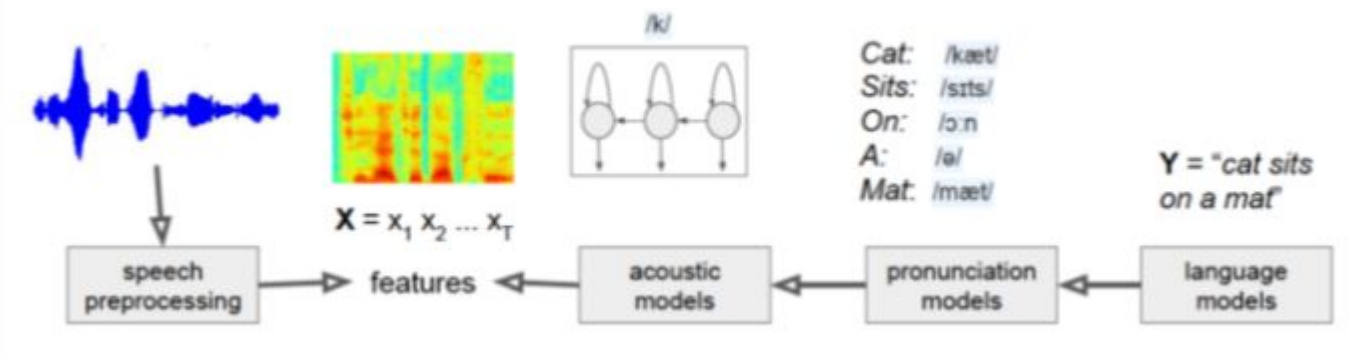
Speech to Text, Text to Speech

Преслав Хаджицанев, ФН 26318 (ИИОЗ)

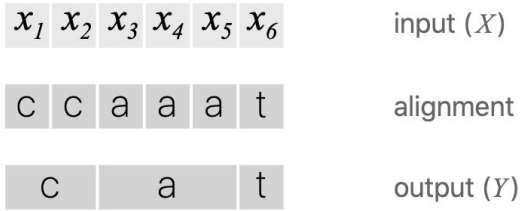
Джовани Чемишанов, ФН 26415 (ИИОЗ)

Speech to text

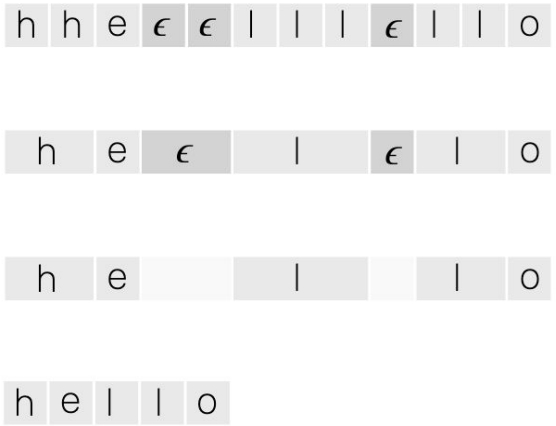


Connectionist Temporal Classification

Интуиция



Подход

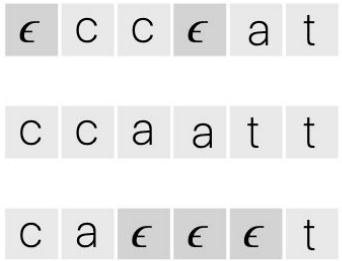


First, merge repeat characters.

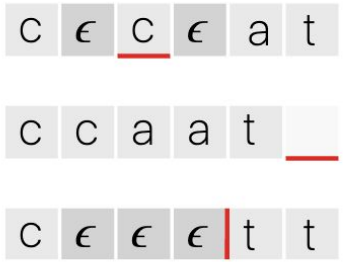
Then, remove any ϵ tokens.

The remaining characters are the output.

Valid Alignments



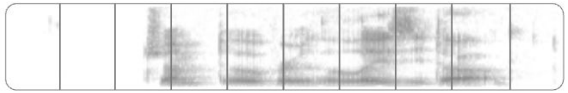
Invalid Alignments



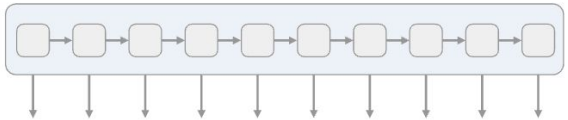
corresponds to $Y = [c, c, a, t]$

has length 5

missing the 'a'



We start with an input sequence, like a spectrogram of audio.



The input is fed into an RNN, for example.

h	h	h	h	h	h	h	h	h	h
e	e	e	e	e	e	e	e	e	e
l	l	l	l	l	l	l	l	l	l
o	o	o	o	o	o	o	o	o	o
ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ

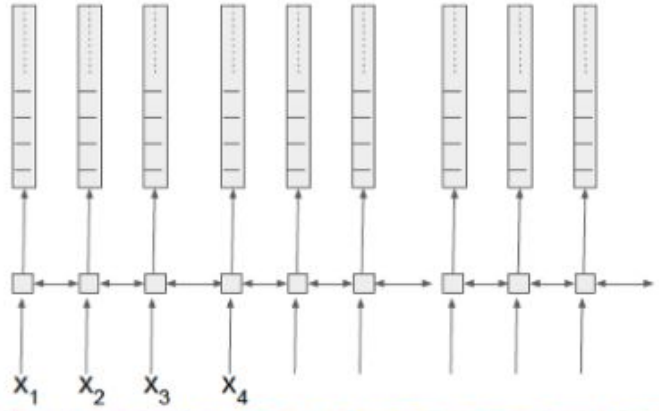
The network gives $p_t(a \mid X)$, a distribution over the outputs {h, e, l, o, ϵ} for each input step.

h	e	ϵ	l	l	ϵ	l	l	o	o
h	h	e	l	l	ϵ	ϵ	l	ϵ	o
ϵ	e	ϵ	l	l	ϵ	ϵ	l	o	o

With the per time-step output distribution, we compute the probability of different sequences

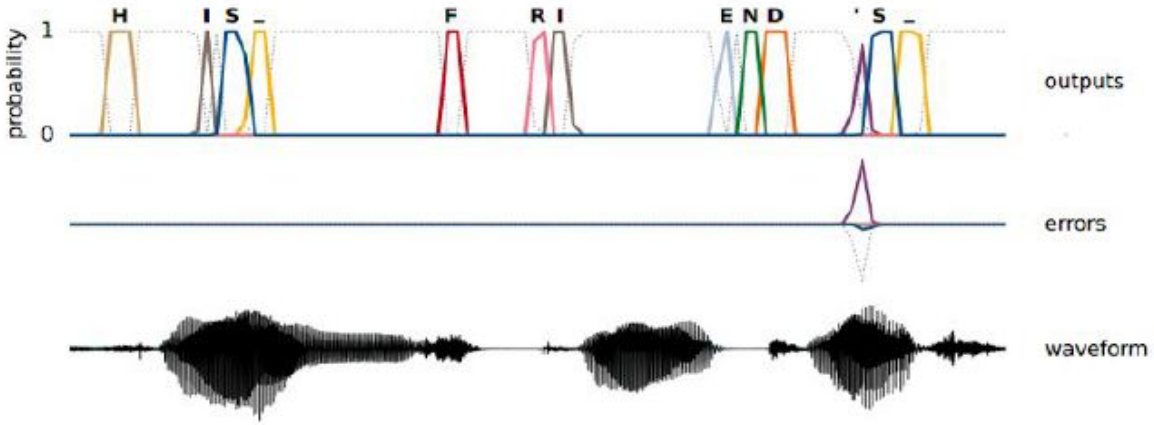
h	e	l	l	o
e	l	l	o	
h	e	l	o	

By marginalizing over alignments, we get a distribution over outputs.

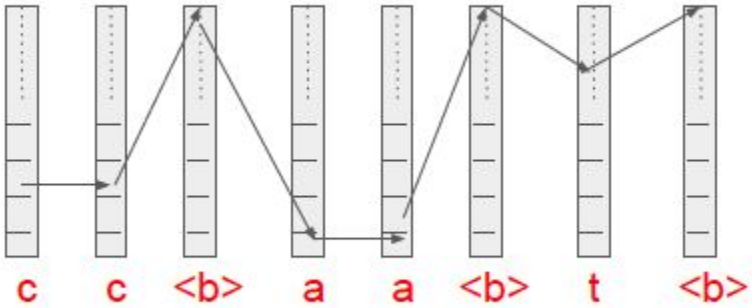


Softmax over vocabulary
{a, b, c, d, e, f, .. z, ?, ., !,..} and extra token <b>.

Softmax at step, t, gives a score s(k,t) = log Pr(k, t|**X**) to category k in the output at time t.



Model learns to make peaky predictions!


c
c
<b>
a
a
<b>
t
<b>

Вероятност за едно разпределение

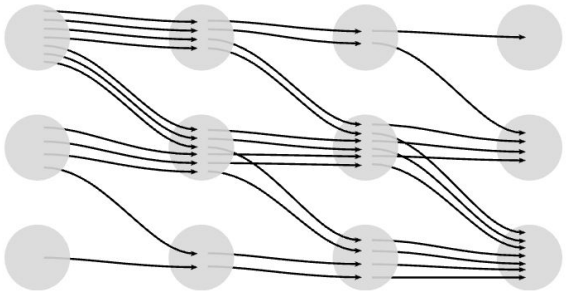
$$p(Y \mid X) = \sum_{A \in \mathcal{A}_{X,Y}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

The CTC conditional **probability**

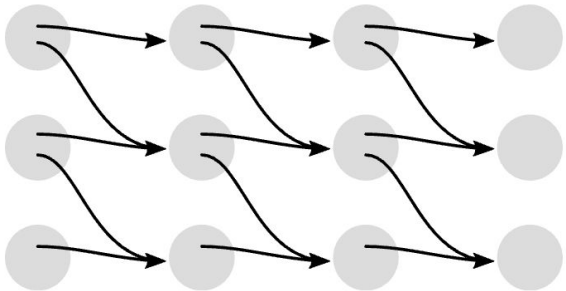
marginalizes over the set of valid alignments

computing the **probability** for a single alignment step-by-step.

Оптимизация



Summing over all alignments can be very expensive.



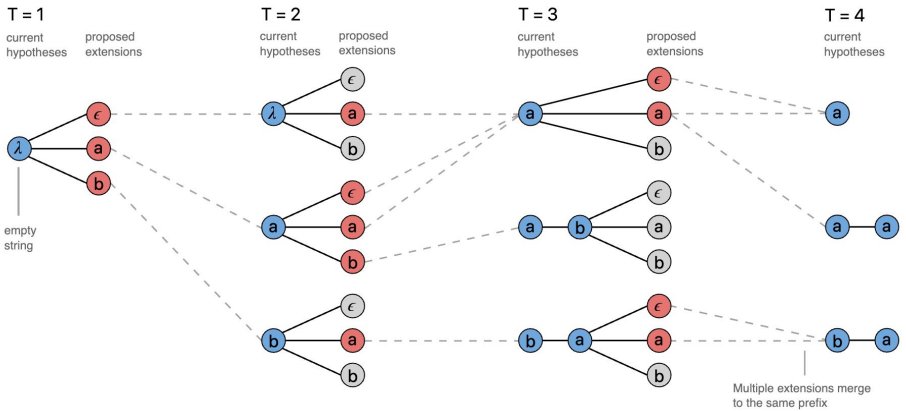
Dynamic programming merges alignments, so it’s much faster.

Вземане на решение

Максимална вероятност

$$A^* = \underset{A}{\operatorname{argmax}} \prod_{t=1}^{T} p_t(a_t \mid X)$$

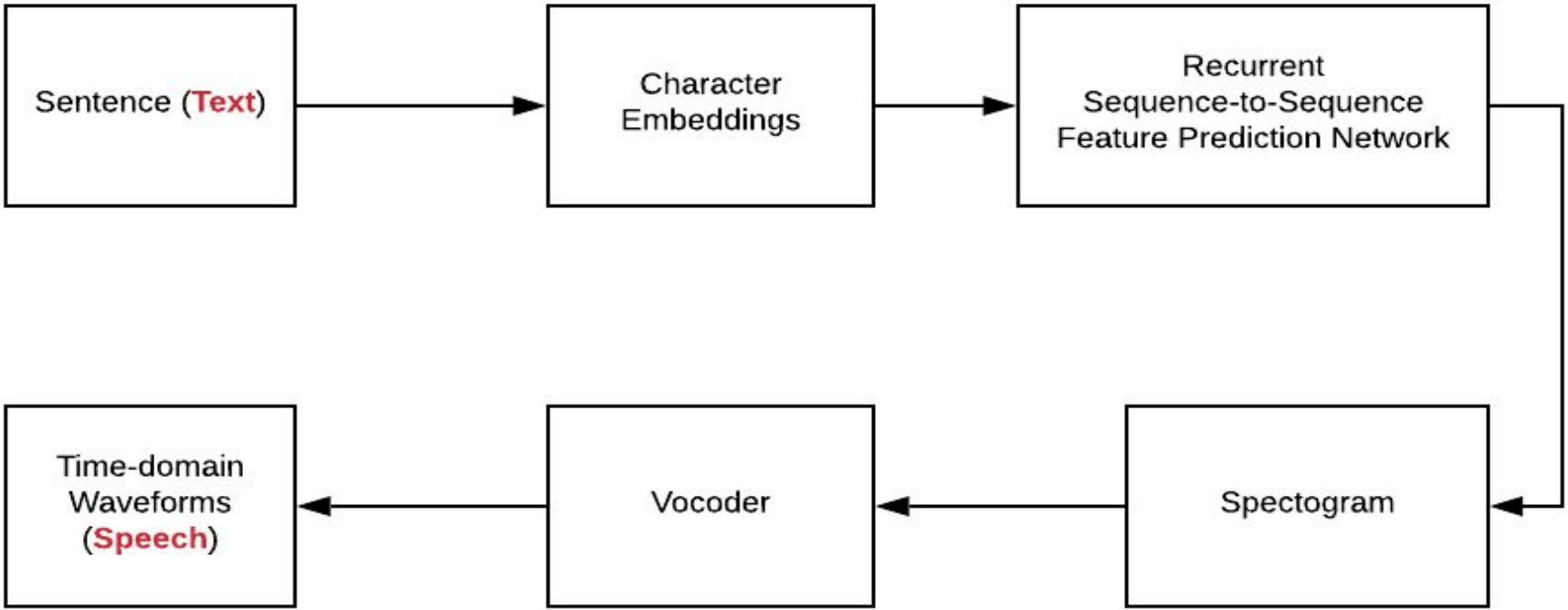
Beam Search



Употреби, Проблеми, Сравнения, Решения

Text to Speech

Подход


Sentence (Text)
Character Embeddings
Recurrent Sequence-to-Sequence Feature Prediction Network
Spectogram
Vocoder
Time-domain Waveforms (Speech)

Deep Voice

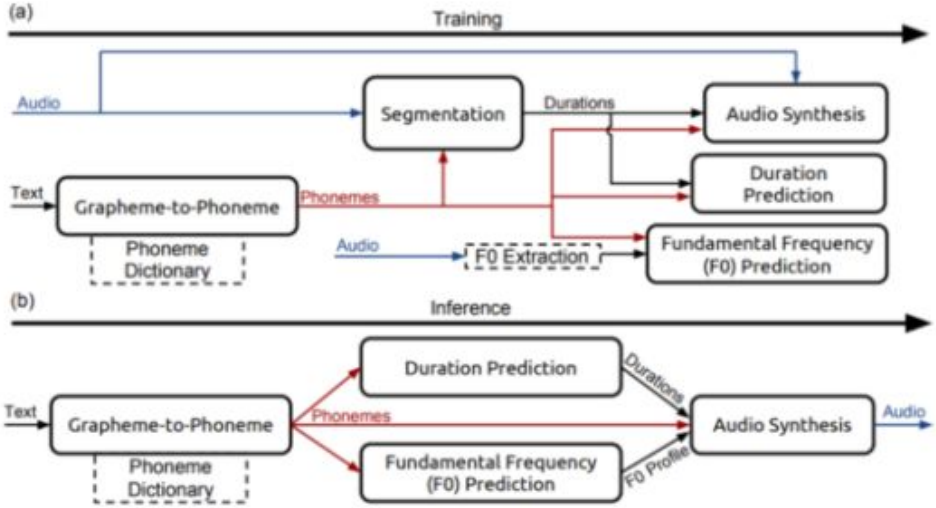


Figure 1. System diagram depicting (a) training procedure and (b) inference procedure, with inputs on the left and outputs on the right. In our system, the duration prediction model and the F0 prediction model are performed by a single neural network trained with a joint loss. The grapheme-to-phoneme model is used as a fallback for words that are not present in a phoneme dictionary, such as CMUDict. Dotted lines denote non-learned components.

Wavenet

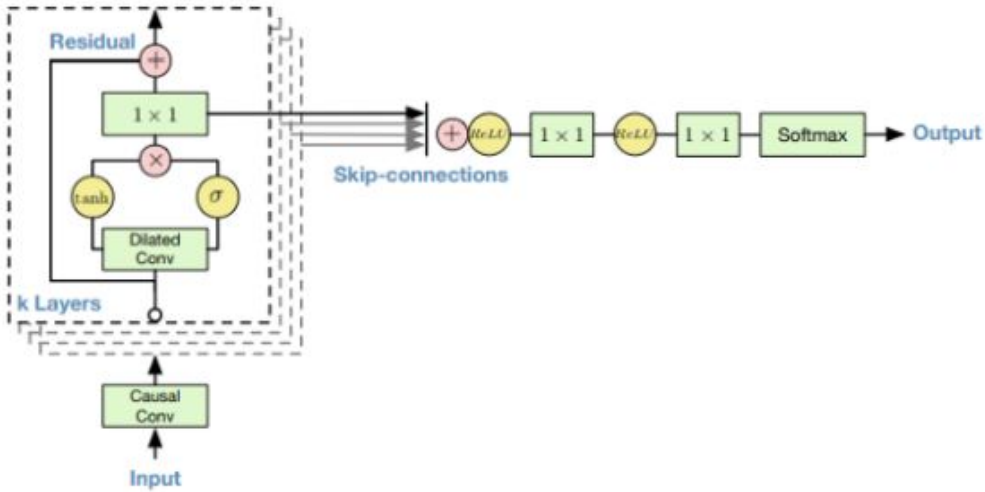
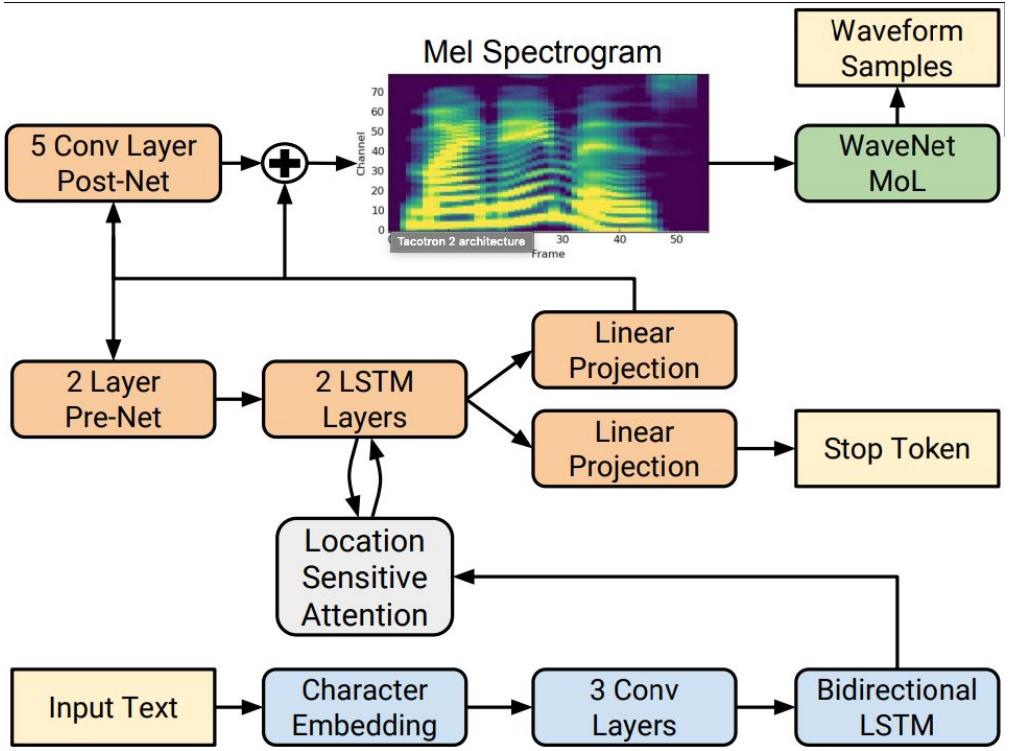


Figure 4: Overview of the residual block and the entire architecture.

Tacotron 2


Mel Spectrogram
Waveform Samples
5 Conv Layer Post-Net
WaveNet MoL
Channel
Frame
Tacotron 2 architecture
2 Layer Pre-Net
2 LSTM Layers
Linear Projection
Linear Projection
Stop Token
Location Sensitive Attention
Input Text
Character Embedding
3 Conv Layers
Bidirectional LSTM

Transformers TTS

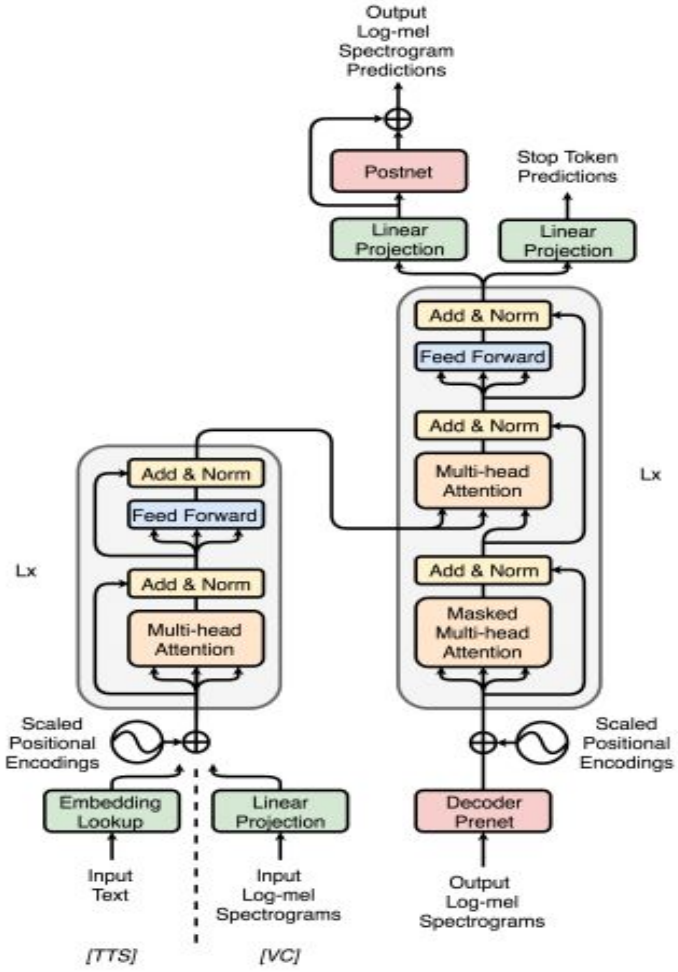


Figure 1: *Model architecture of Transformer-TTS and VC.*

Demo page Microsoft
https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

What is the best text to speech API?

After reviewing all the text to speech APIs, we found these 10 APIs to be the very best and worth mentioning:

- IBM Watson API
- Rev.ai API
- Speechmatics API
- Google Speech-to-text API
- Robomatic.ai API
- Amazon Polly API
- Voicepods API
- Dialog Flow API
- Microsoft Azure Cognitive Services API
- Ispeech API