ITMO.Hack

ITMO ML LAB

BOTAN INVESTMENTS

SOTA ML School

SOTA in Text To Speech

Евгений Шуранов
Huawei Principal Engineer
ITMO associate professor
PhD
evgeniy.shuranov@huawei.com

ITMO UNIVERSITY

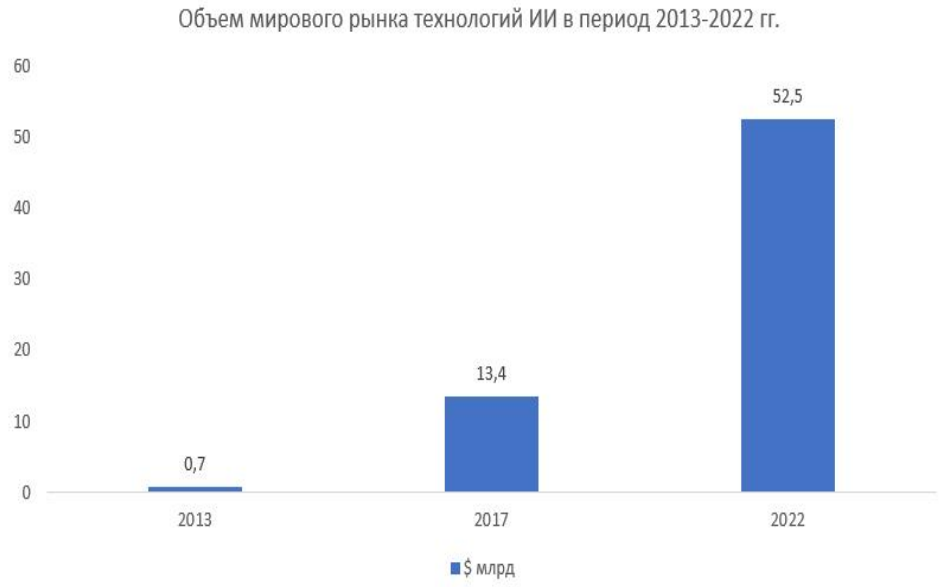
HUAWEI

Agenda

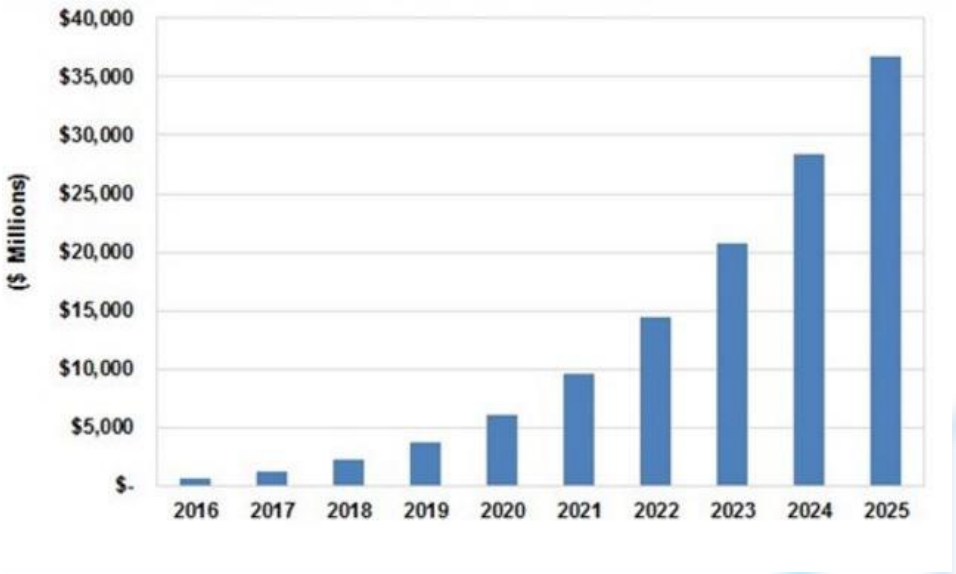
- Overview
  - Directions in speech processing
  - First steps
- SOTA TTS
- Example of interesting research in TTS
  - Multi-lingvo Multi-speaker Text-to-Speech
  - TTS improves speech recognition
  - Emotional Text-to-Speech system
- About our team and contacts

01

Рост рынка ИИ

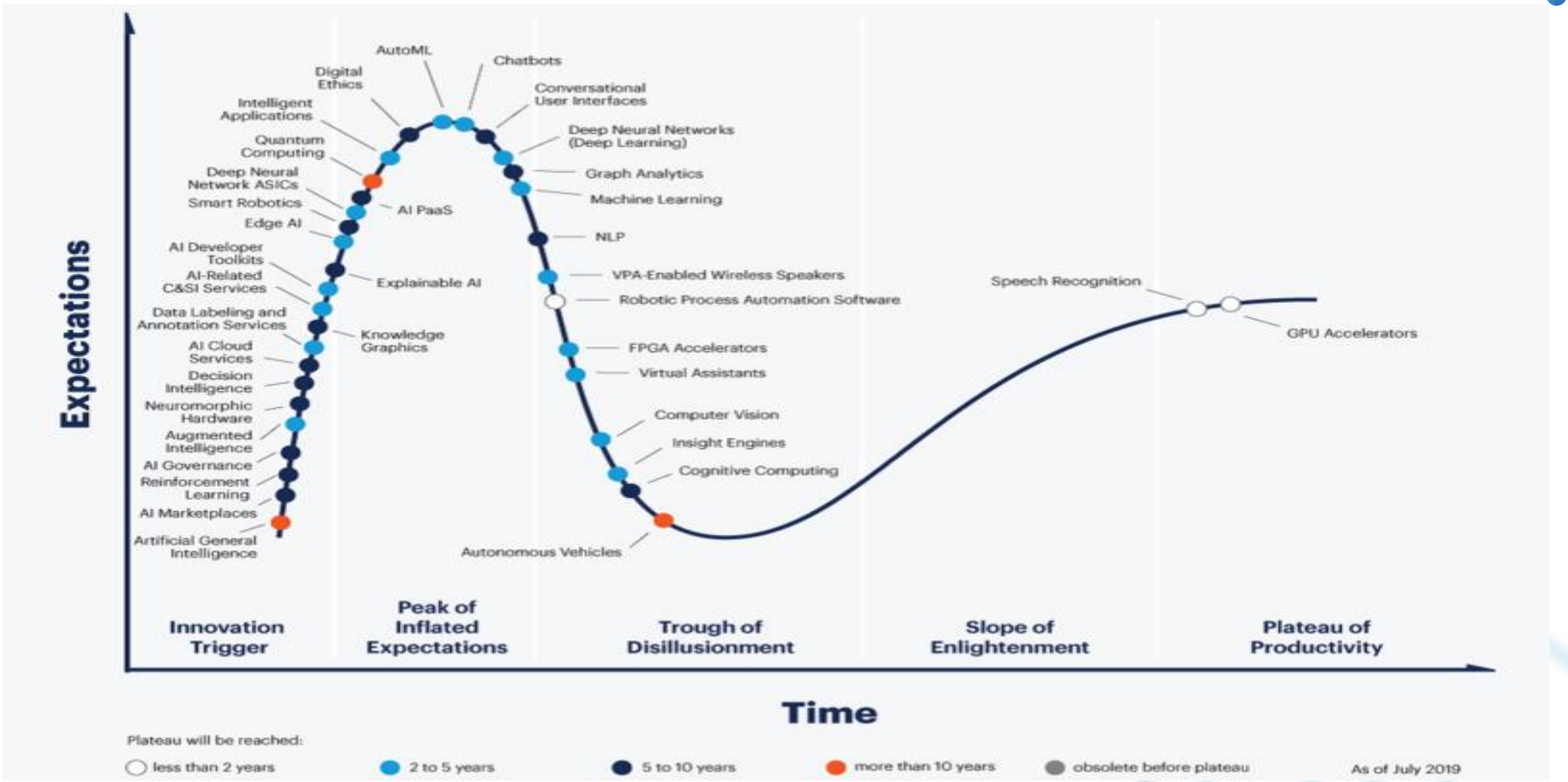


Доходы рынка ИИ 2016-2015. Данные Tractica



Рынок ИИ будет расти на 31% ежегодно -Frost & Sullivan

Expectations



How to start:

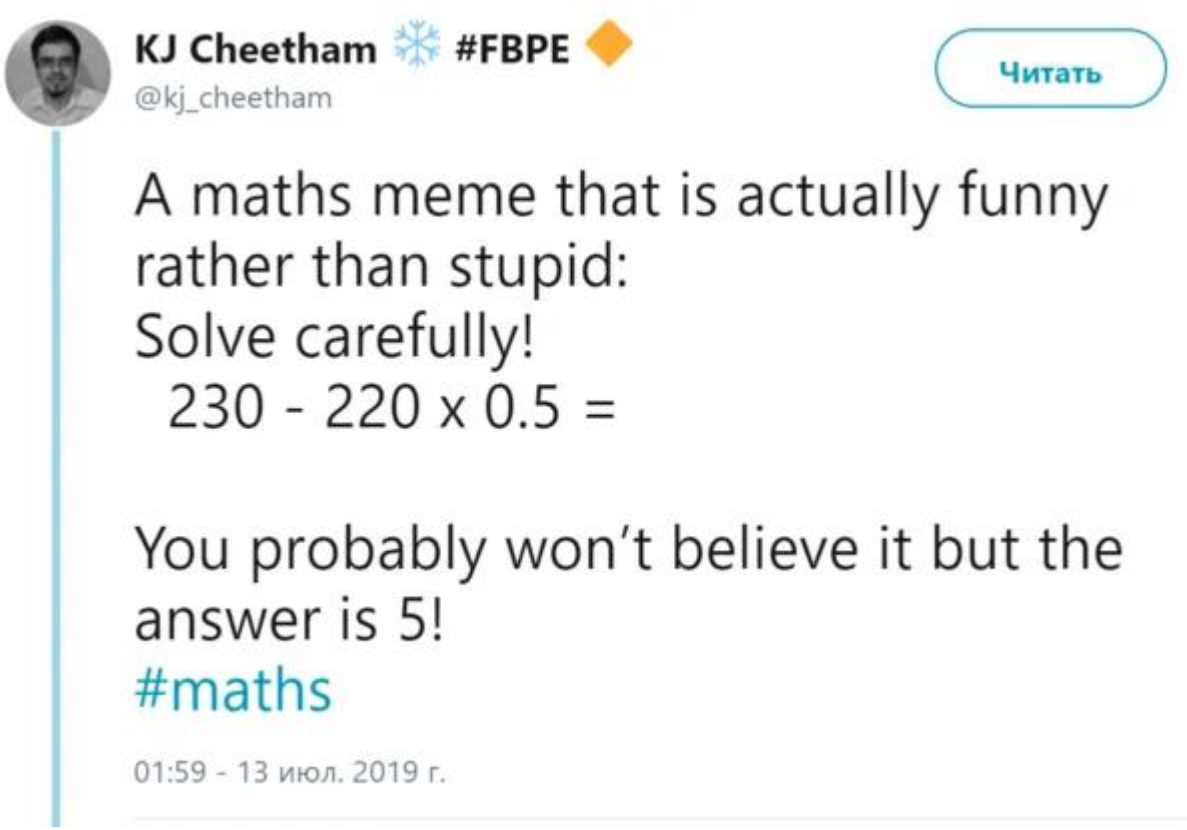
What you need	How to find
Papers	Arxiv / conferences /**challenges** /habr /telegram /slark and e.t.c.
Source examples	----------
Datasets	----------
GPU	google colab / universities

Main events in speech technology

Name	Main domain	Type
Blizzard	TTS	Chalenge
Voice Conversion Challenge	TTS / Voice Conversion	Chalenge
Chime	ASR	Chalenge
NIST	Voice recognition	Chalenge
Interspeech	ASR++	Conference
Speecom	ASR++	Conference
Icasp	ASR++	Conference
EmotiW	Emo recognition	Chalenge
Dcase	Event detection	Chalenge
OMG	Emo recognition	Chalenge

First step errors

1 Math is always not enough!



2 If you good at math remember about physics! (or business value ;))

What is TTS? Applications?

Any text to produce!

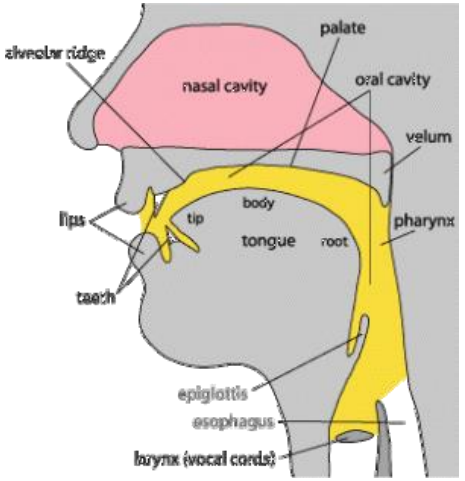
⇩

Text-to-Speech method

⇩

Intelligible speech

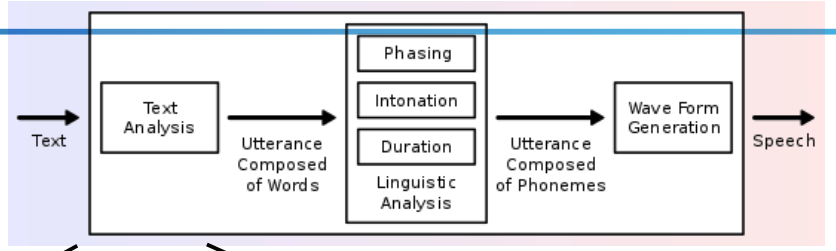
In 1779 the German scientist Christian Gottlieb Kratzenstein won the first prize in a competition announced by the Russian Imperial Academy of Sciences and Arts for models he built of the human vocal tract that could produce the five long vowel sounds. In International Phonetic Alphabet notation:
[aː], [eː], [iː], [oː] and [uː].



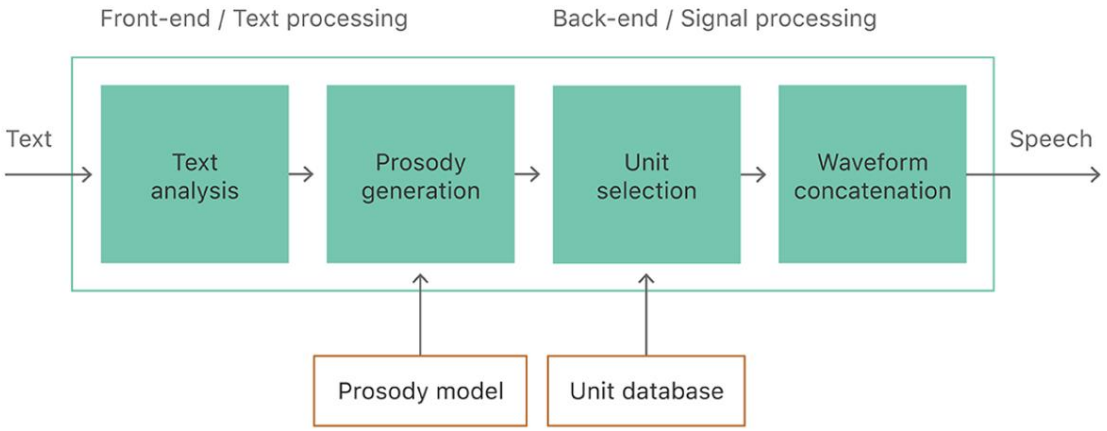
Conventional TTS technologies

[7]

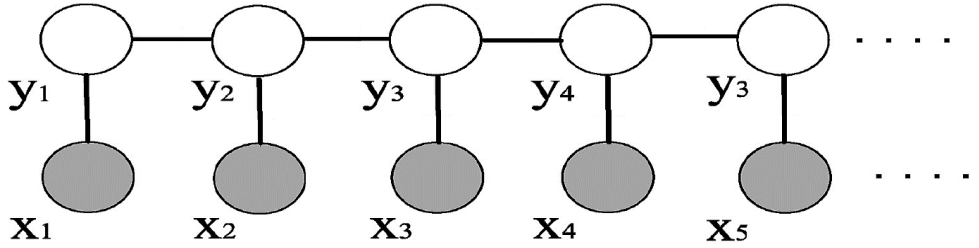
Overview of a typical TTS system



Concatenative unit selection



Statistical parametric Synthesis



Viterbi algorithm to produce speech waveform

Conventional TTS technologies

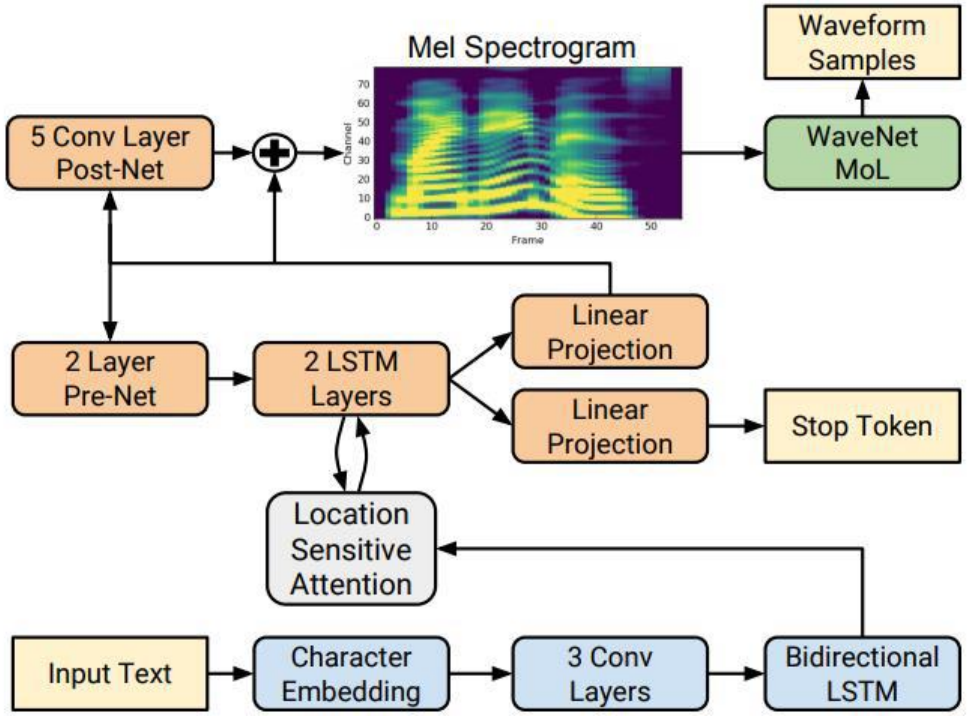
[8]

Front-End: Text preprocessing

1. **Normalization**
   1. She was born on 1994
   2. She was born on nineteen ninety four
2. **Phonemization** (IPA or Arpabet for English)
   1. She was born on 1994 –
   2. SH IY1 / W AA1 Z / B AO1 R N / AA1 N / W AH1 N / TH AW1 . Z AH0 N D / N AY1 N / HH AH1 N . D R AH0 D / AH0 N D / N AY1 N . T IY0 / F AO1 R
   3. 你好， 你好吗？(nǐ hǎo nǐ hǎo mǎ)

	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CIA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0·6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3·20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3·45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3·45 billion*
	PRCT	percentage	*75%, 3·4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car* (see also SLNT and PUNC examples)
	SLNT	not spoken, word boundary or emphasis character:	
M		word boundary	*M.bath, KENT*RLTY, _really_*
I	PUNC	not spoken, non-standard punctuation: "***" in	
S		phrase boundary	*$99,9K***Whites*, "..." in *DECIDE...Year*
C	FNSP	funny spelling	*sllooooooww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

SOTA TTS



System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

- High quality speech
- Meets hardware constraints
- Flexible architecture

https://habr.com/ru/post/465941/

SOTA. Vocoder evolution.

Vocoder	Year	Type of network	Speed	Training time
Wavenet	2016	Fully-convolutional	Extremelyslow (Open versions take several minutes to produce 1 second)	Weeks
WaveRNN	2018	Fullyrecurrent	4xfaster than realtimeon GPU	Around 1 week (according to open implementations)
LPCNet	2018	Convolutional+Reccurent+ Algorithmic(LPC)	4x faster than realtimeon CPU	1-2 days
MelGAN	2019	Fully-convolutional	20x faster than realtimeon CPU	3-4days

Frontend

- Languages are very different

Hello Sota Machine learning School:


English


German(on English word)


Russian (on English word)


Russian

Disentaglement

Speech data contains a lot of information:

Speech data = Text+ Speaker+ Prosody+ Recording conditions

Some of theme are present in the annotation and some are not.

-random or flat prosody,

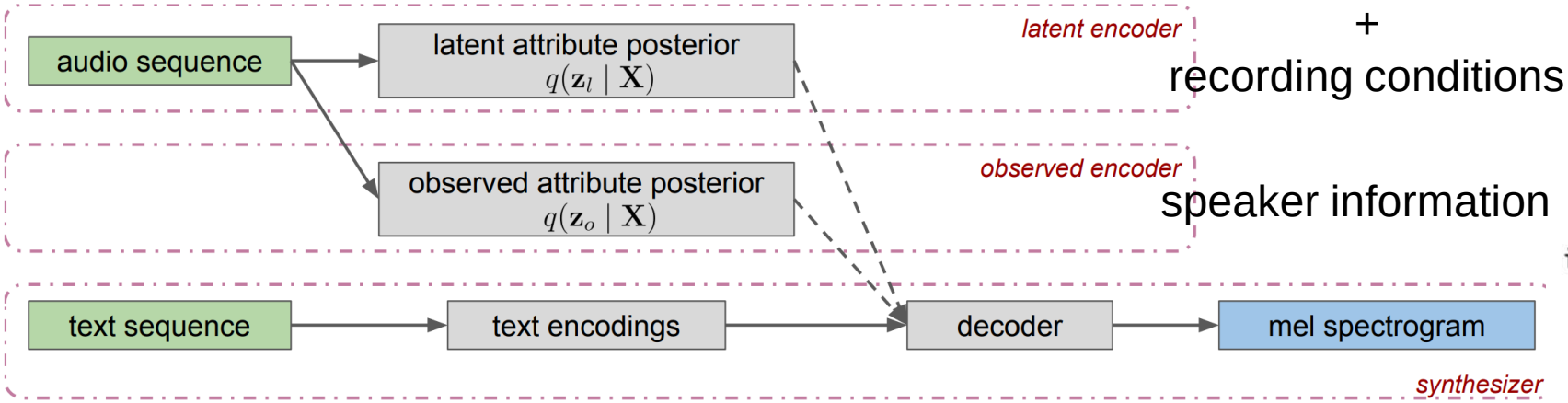
-some level of noise during synthesis if recording artifacts were present in the data.

Aim:

1.We want to control all components of the speech during synthesis.

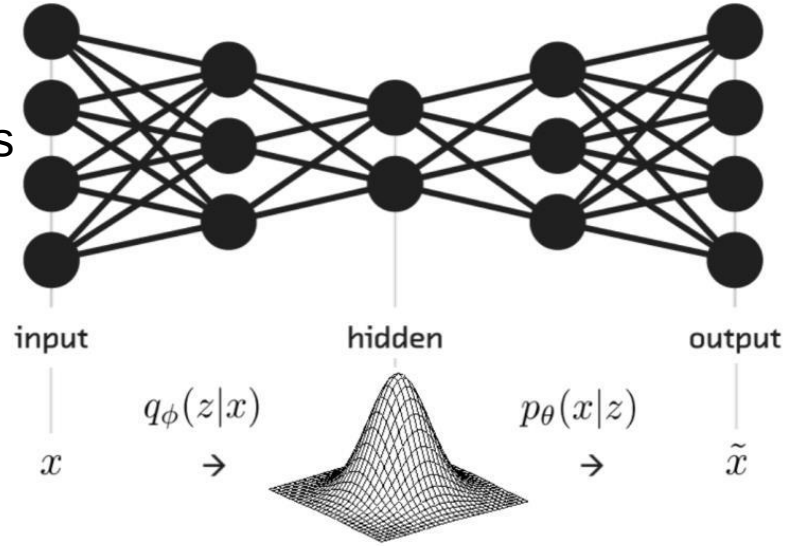
2.We want to train TTS without additional mark-up.

Disentaglement



prosody
+
recording conditions

speaker information



GMVAE:
- uses hierarchical model for latent variables
- most of the component has interpretable role (noise/speed/pitch/accent etc) –easy control

VAE:
- Latent space is not interpretable:
Separate dimensions of the vector have no meaning
- Due to the above control (in practice) is problematic

arxiv.org/pdf/1810.07217.pdf

Multi-lingvo Multi-speaker Text-to-Speech

Target:

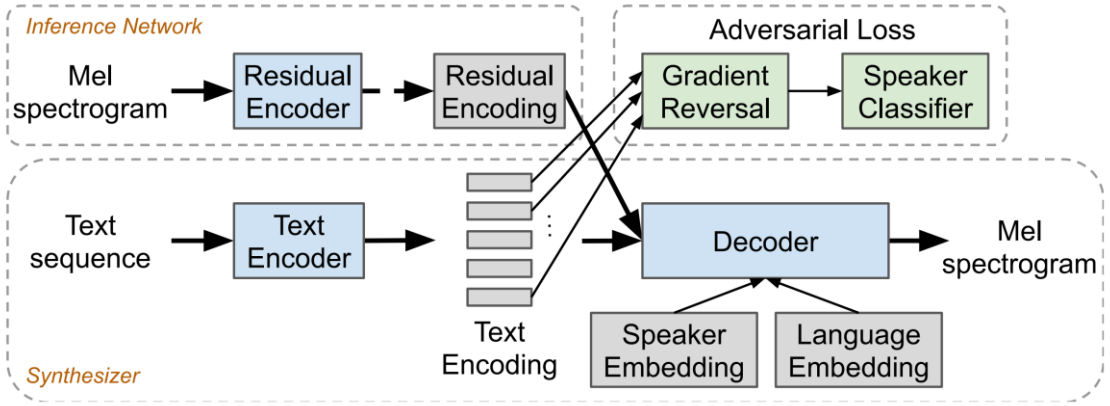
Complexity: <50% bigger that single speaker model
Size: <50% bigger that single speaker model
Quality: same MOS as single speaker model

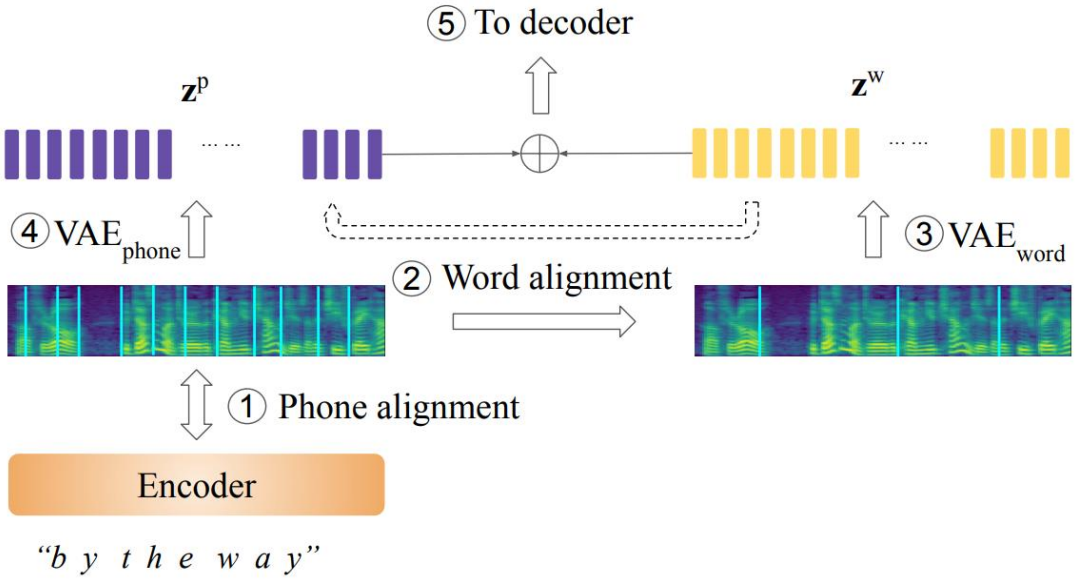
Challenges:

Speaker/accent/language disentanglement
NO product-ready open technology

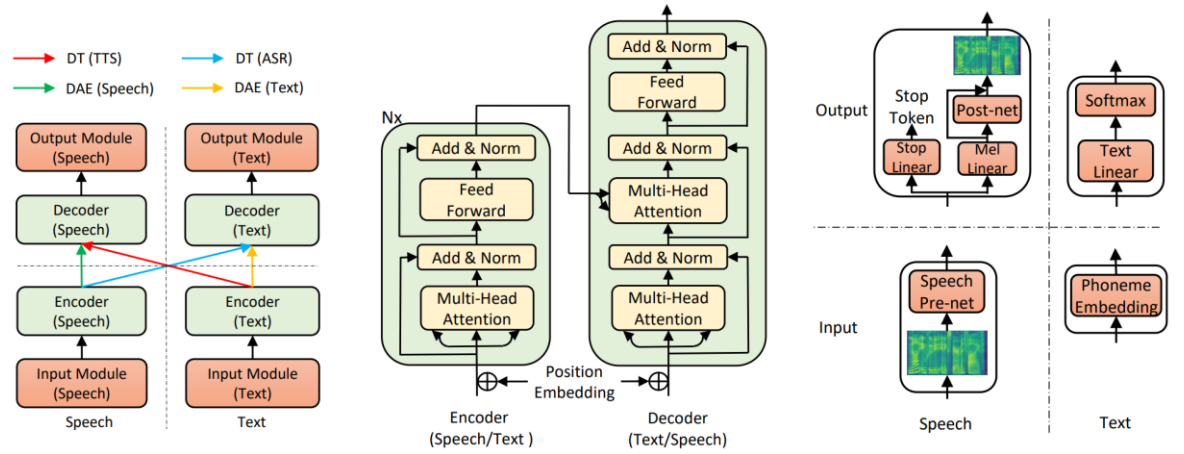
Key technology:

VAE + RNN combination





TTS improves speech recognition



(a) Unified training flow (b) Encoder/Decoder for speech/text (c) Input/Output module for speech/text

Figure 1. The overall model structure for TTS and ASR. Figure (a): The unified training flow of our method, which consists of a denoising auto-encoder (DAE) of speech and text, and dual transformation (DT) of TTS and ASR, both with bidirectional sequence modeling. Figure (b): The speech and text encoder and decoder based on Transformer. Figure (c): The input and output module for speech and text.

Method	MOS (TTS)	PER (ASR)
GT	4.54	-
GT (Griffin-Lim)	3.21	-
Supervised	3.04	2.5%
Pair 200	Null	72.3%
Our Method	2.68	11.7%

Table 1. The comparison between our method and other systems on the performance of TTS and ASR.

Paired Data	100	200	300	400	500
PER (ASR)	64.2%	11.7%	8.4%	5.2%	4.4%
MOS (TTS)	Null	2.45	2.49	2.64	2.78

Table 3. The PER on ASR with different amount of paired data for our method.

Authors used mutual training of unified transformers architecture for TTS, ASR.

Authors achieved 99.84% in terms of word level intelligible rate and 2.68 MOS for TTS, and **4.4% PER for ASR with just 500 paired data on LJSpeech** dataset. They stated to prove that it is possible to train ASR with close to SOTA performance only with 500 audio clips (~1 hour).

Emotional Text-to-Speech system

1) *Speech Emotion Recognition (2020-2021)*
   - *by voice*
   - *by text*
2) *Emotional Text-to-Speech System (2020-2022)*

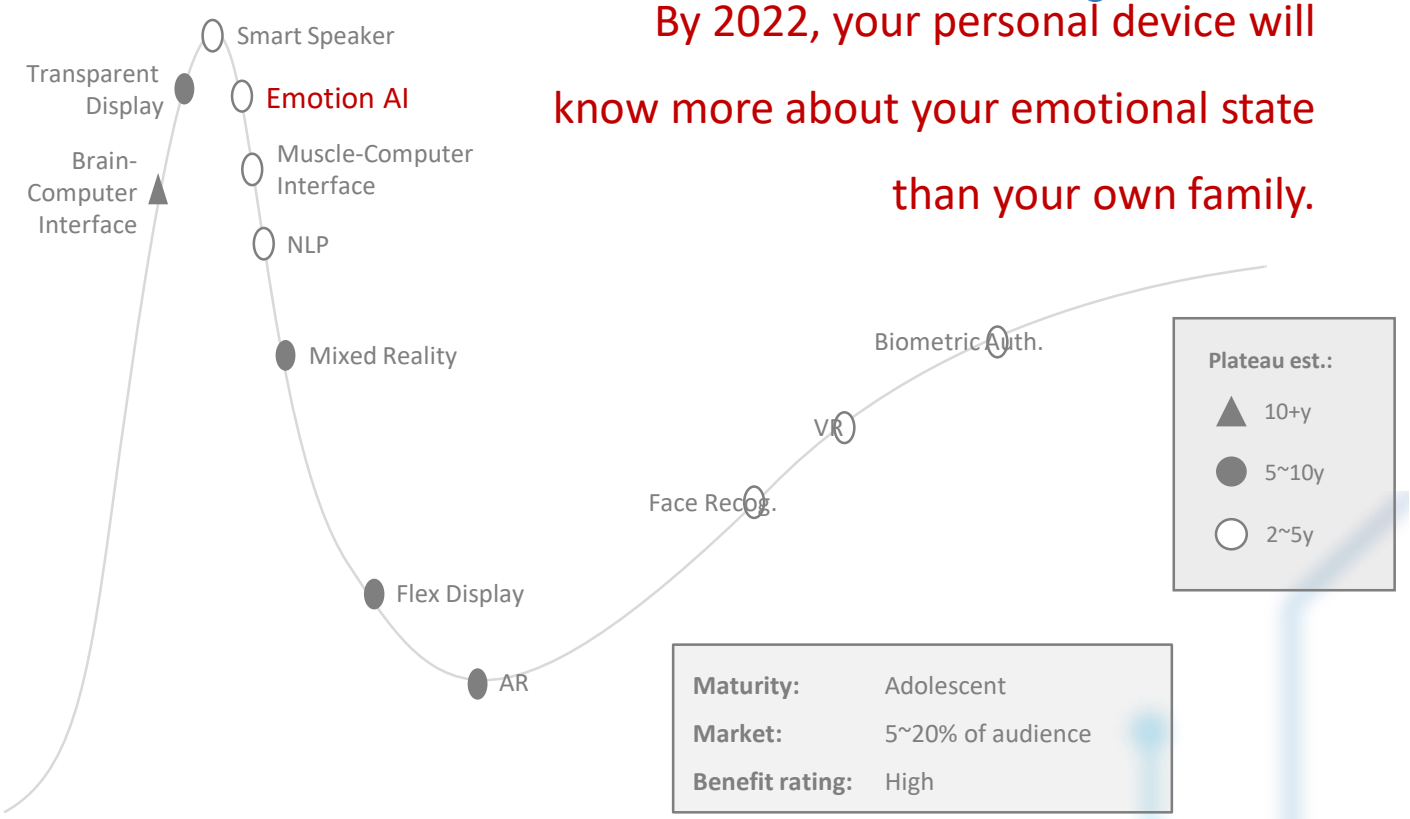
Challenges:

poorly defined task
- not determined basic emotion list

dataset transfer problems
- language transfer
- audio/text/video transfer
- actors/spontaneous speach transfer
- different labeling process

model robustness and stability

NO product-ready open technology



By 2022, your personal device will know more about your emotional state than your own family.

Key technology:

pretrained features + LSTM/CNN combination

Own research

TTS improvement

Controlled TTS

Voice cloning

Accent transfer

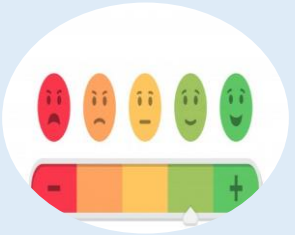
ASR improvement

on device

Event detection

Punctuation

Cooperation research



Emotion recognition (Voice)

Emotional TTS

Emotion recognition (Text)

Voice-print technology

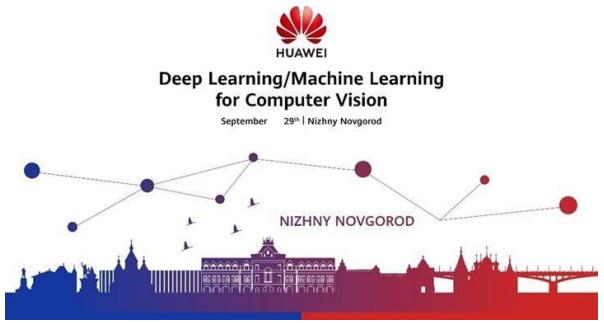
Anti spoofing

GANs approach (ASR + TTS)

Goals of my team

- Find new perspective research directions in the speech domain
- Achieve result better then SOTA and find ways to implement this result for CBG products
- Make POC and production ready prototype
- Support delivery process of technology to production team

Workshops & open days 2020 in SPb



01

ITMO UNIVERSITY

СПбГУ

ПОЛИТЕХ

Thank you for your attention

Welcome to cooperation!

01 https://docs.google.com/forms/d/e/1FAIpQLSfXzHrmdxdOKizMksoFi3IMSPoG_JMz4FH6QWYJ43IgUidIQA/viewform

Contacts

Cooperation: viktor.rakhmanov@huawei.com HR: nikulina.daria@huawei.com