




Russian-language speech recognition system based on DeepSpeech

Oleg Jakushkin, George Fedoseev,
Anna Shaleva, Olga Sedova

Why?



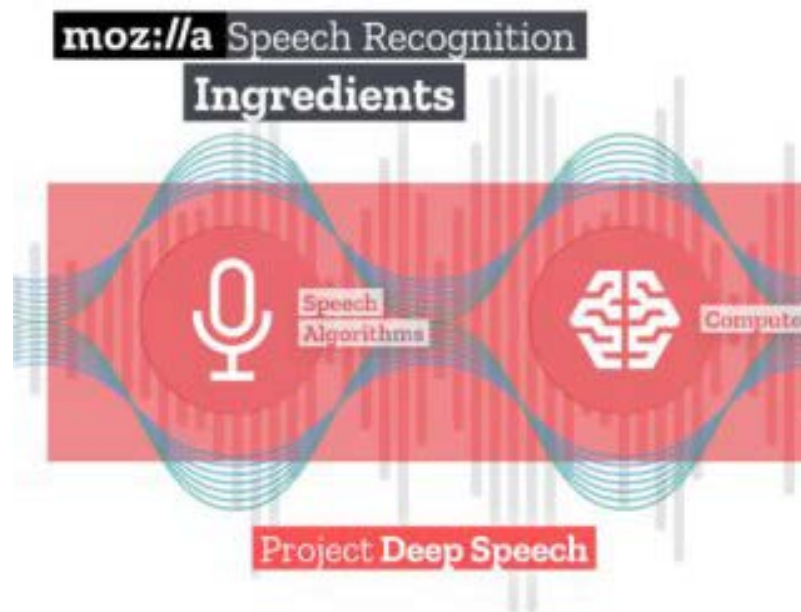
- create better recognition system for Russian speech using modern technologies (Deep Learning, End-to-End speech recognition)
- 

Why?




Sphinx

 KALDI





How to measure recognition precision?

Word Error Rate (WER)

Word error rate can be computed as:

$$WER = \frac{S + D + I}{N}$$

where

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of words in transcript

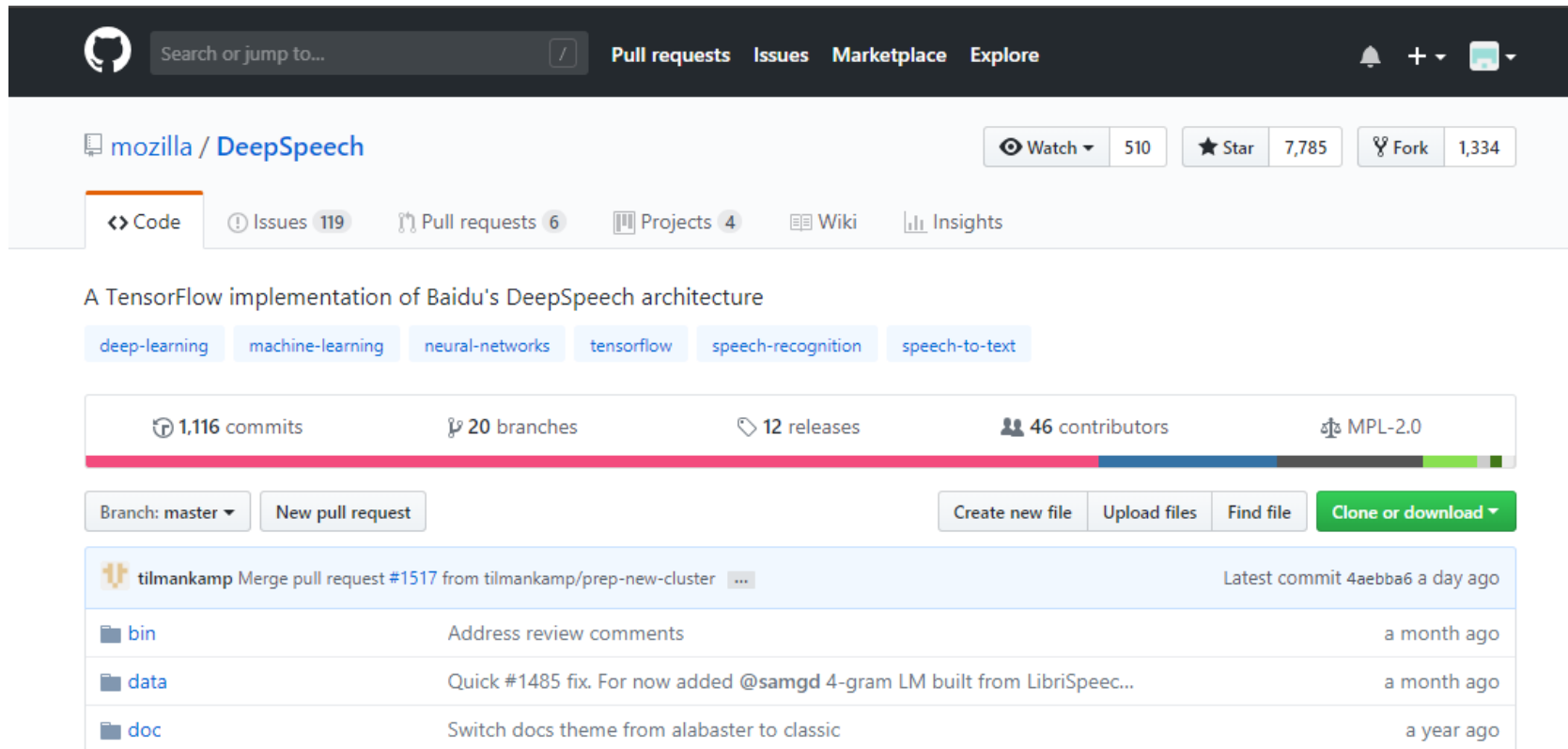
(1 substitution + 1 deletion) / 6 = 0.333

```
WER: 0.333333, loss: 21.334461, mean edit distance: 0.161290
- src: " вы видели откуда я взял данные"
- res: " вы видели откуда я вделдарны"
-----
WER: 0.400000, loss: 8.423505, mean edit distance: 0.107143
- src: " однажды в город пришла беда"
- res: " однажды в город пришлабита"
```

(1 substitution + 1 deletion) / 5 = 0.4

Where to start?

Adapt Open Source implementation with trained English model from Mozilla...



The screenshot shows the GitHub repository page for mozilla/DeepSpeech. The repository is a TensorFlow implementation of Baidu's DeepSpeech architecture. It has 510 watchers, 7,785 stars, and 1,334 forks. The repository is licensed under MPL-2.0 and has 1,116 commits, 20 branches, 12 releases, and 46 contributors. The repository is currently on the master branch. The commit history shows a recent merge pull request #1517 from tilmankamp/prep-new-cluster, and three recent commits: bin (Address review comments, a month ago), data (Quick #1485 fix. For now added @samgd 4-gram LM built from LibriSpeec..., a month ago), and doc (Switch docs theme from alabaster to classic, a year ago).

Search or jump to... / Pull requests Issues Marketplace Explore

mozilla / DeepSpeech Watch 510 Star 7,785 Fork 1,334

Code Issues 119 Pull requests 6 Projects 4 Wiki Insights

A TensorFlow implementation of Baidu's DeepSpeech architecture

deep-learning machine-learning neural-networks tensorflow speech-recognition speech-to-text

1,116 commits 20 branches 12 releases 46 contributors MPL-2.0

Branch: master New pull request Create new file Upload files Find file Clone or download

tilmankamp Merge pull request #1517 from tilmankamp/prep-new-cluster Latest commit 4aebba6 a day ago

bin	Address review comments	a month ago
data	Quick #1485 fix. For now added @samgd 4-gram LM built from LibriSpeec...	a month ago
doc	Switch docs theme from alabaster to classic	a year ago

Where to start?

... with promising results:

A Journey to <10% Word Error Rate



By [Reuben Morais](#)

Posted on November 29, 2017 in [Featured Article](#) and [Research](#) ♥ Share This

At Mozilla, we believe speech interfaces will be a big part of how people interact with their devices in the future. Today we are [excited to announce](#) the initial release of our [open source speech recognition model](#) so that anyone can develop compelling speech experiences.

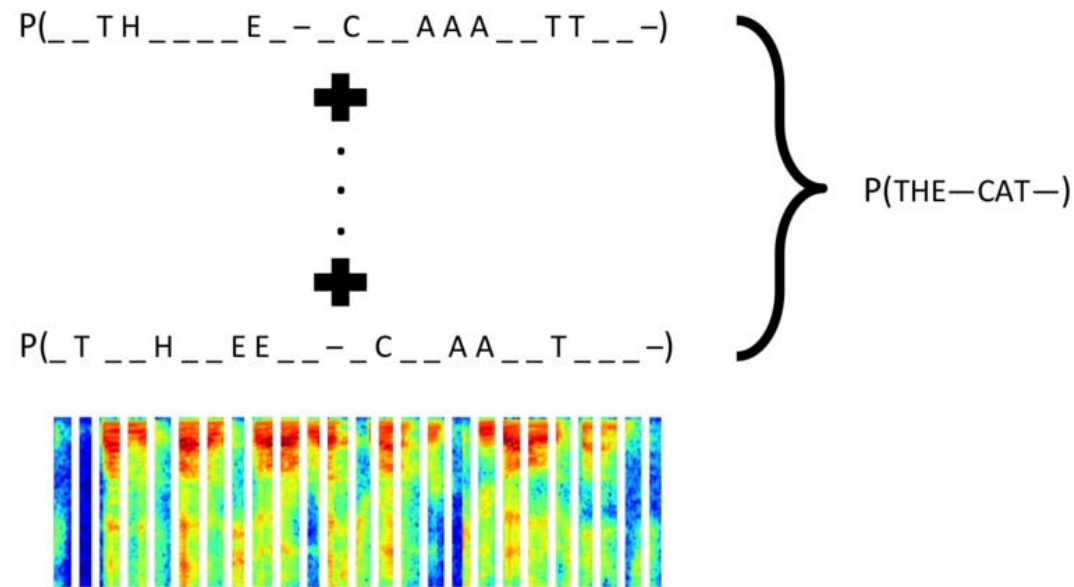
The Machine Learning team at Mozilla Research has been working on an open source Automatic Speech Recognition engine modeled after the Deep Speech papers ([1](#), [2](#)) published by Baidu. One of the major goals from the beginning was to achieve a Word Error Rate in the transcriptions of under 10%. We have made great progress: [Our word error rate on LibriSpeech's test-clean set is 6.5%](#), which not only achieves our initial goal, but gets us close to human level performance.

This post is an overview of the team's efforts and ends with a more detailed



How it works?

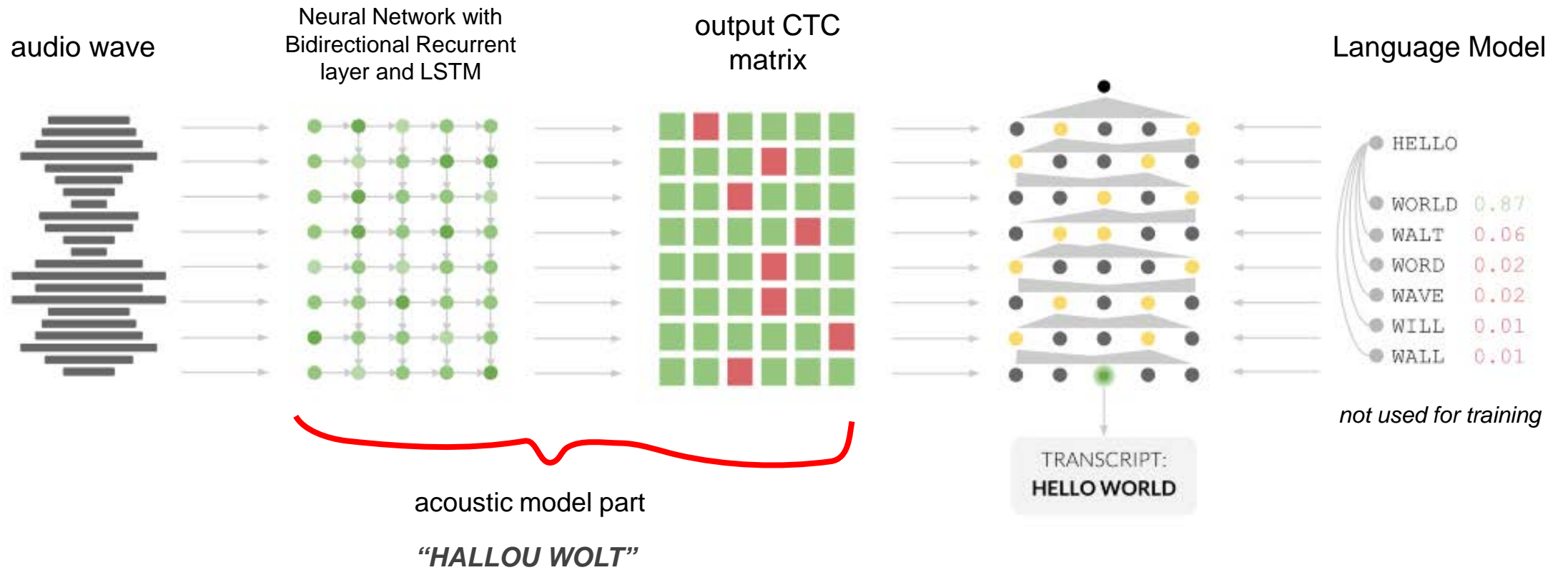
Connectionist Temporal Classification (CTC) approach



Sound \rightarrow Letter alignment
independent approach to
training Recurrent Neural
Network

Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.

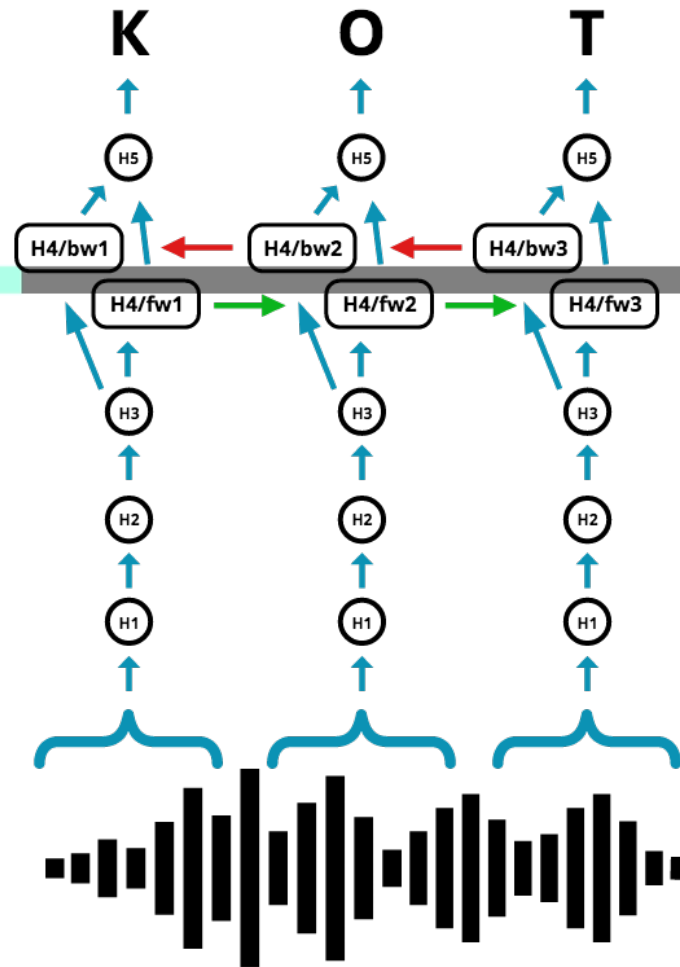
Turning audio into text pipeline



Acoustic model's Neural Network structure



- 5 hidden layers (pretty deep)
- 4th layer: BiRNN with LSTM cells



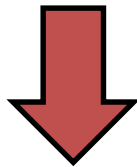
How to train acoustic model's neural network?



pairs (audio, transcript)



certain neural network structure



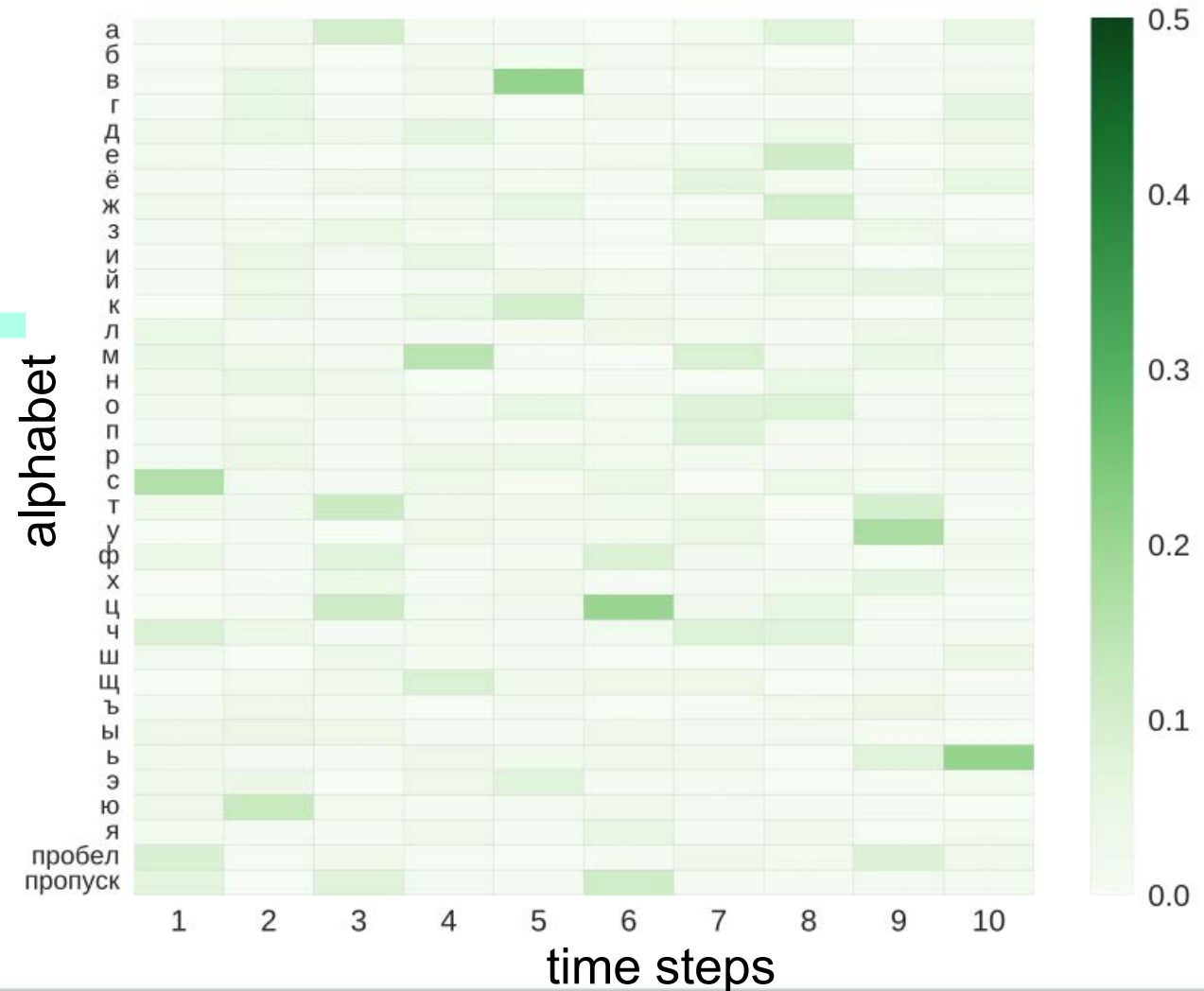
loss function for optimization

What do we get from acoustic model?



CTC output matrix (last NN layer)

each column -
probability distribution over alphabet
symbols for time t



What do we get from acoustic model?

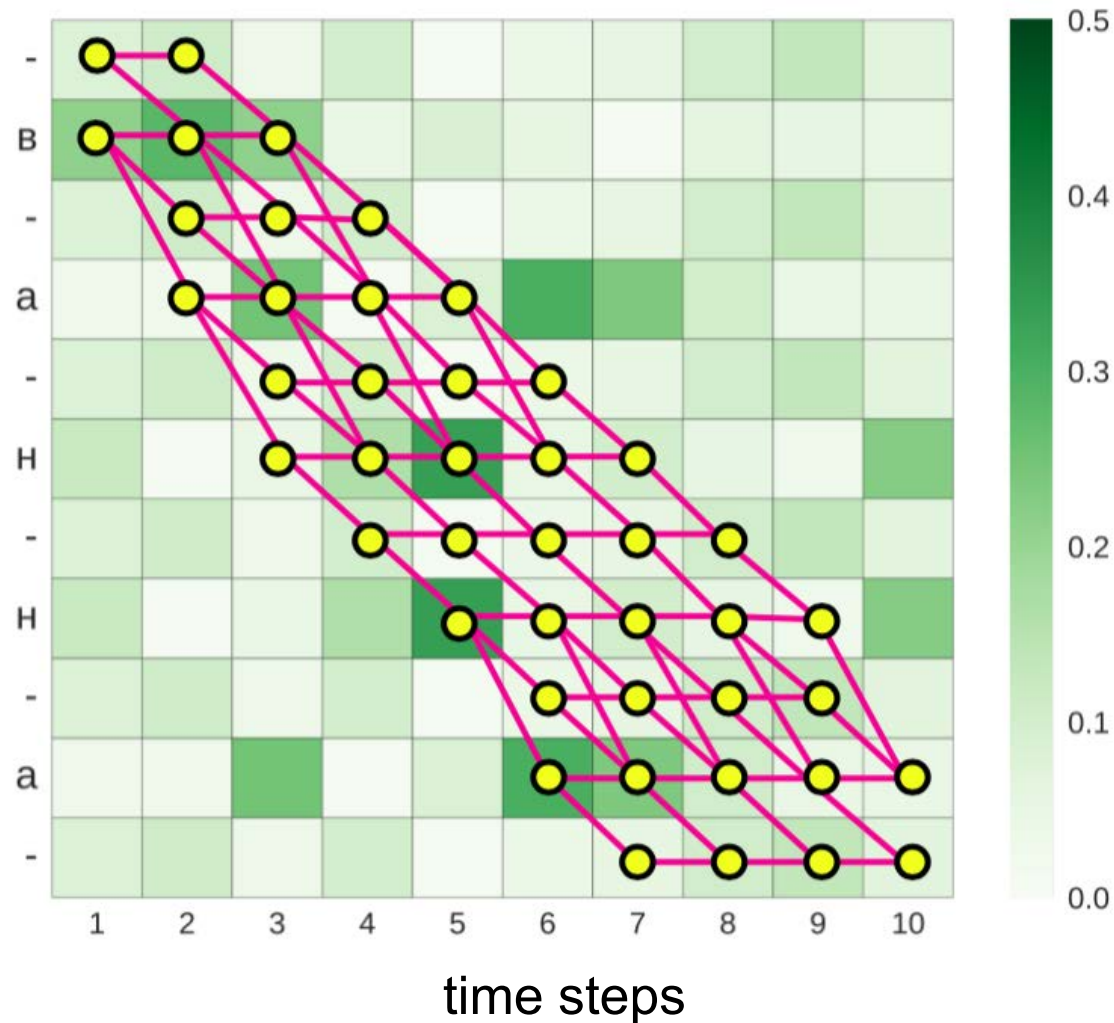


Everyone speaks with different speed

All possible paths corresponding to one labelling “ванна”

For example:

- В-а-Н-на
- В--ааН-на
- В-аааН-на-
- В-аНН-на-
- ВВАН--на-



Optimizing CTC Loss calculation



CTC Loss Calculation Algorithms Comparison



16% speed improvement

when using Baidu's CTC loss calculation implementation

Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." *International Conference on Machine Learning*. 2016.



How to decode trained network output?

- **greedy (max) decoding**
- **prefix search**
- **prefix search with LM**

How to decode trained network output?



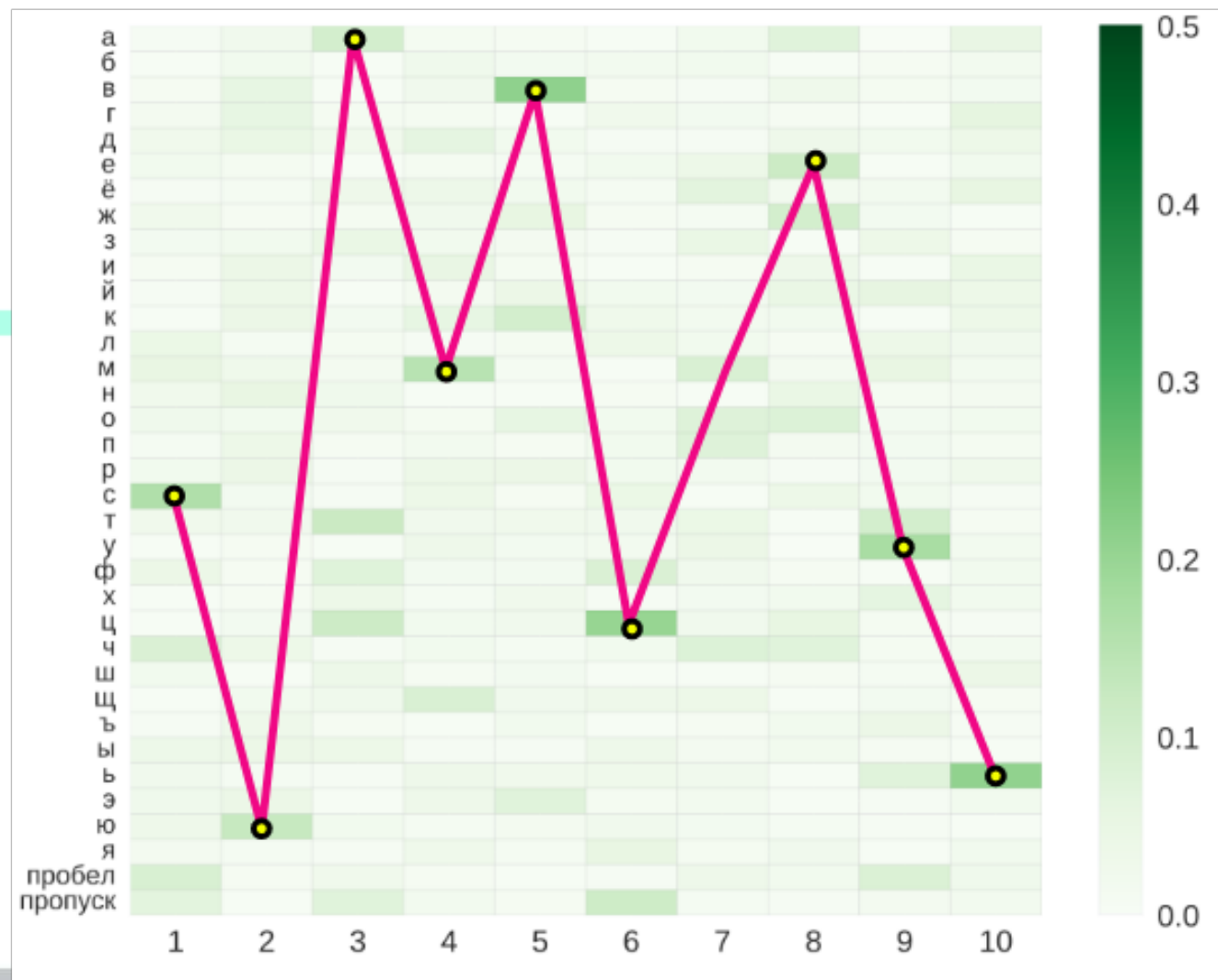
Greedy (max) decoding

---B-OO--XXX-__--BBUUNN-NI---

-B-O-X-_-BUN-NI->

BOX_BUNNI>

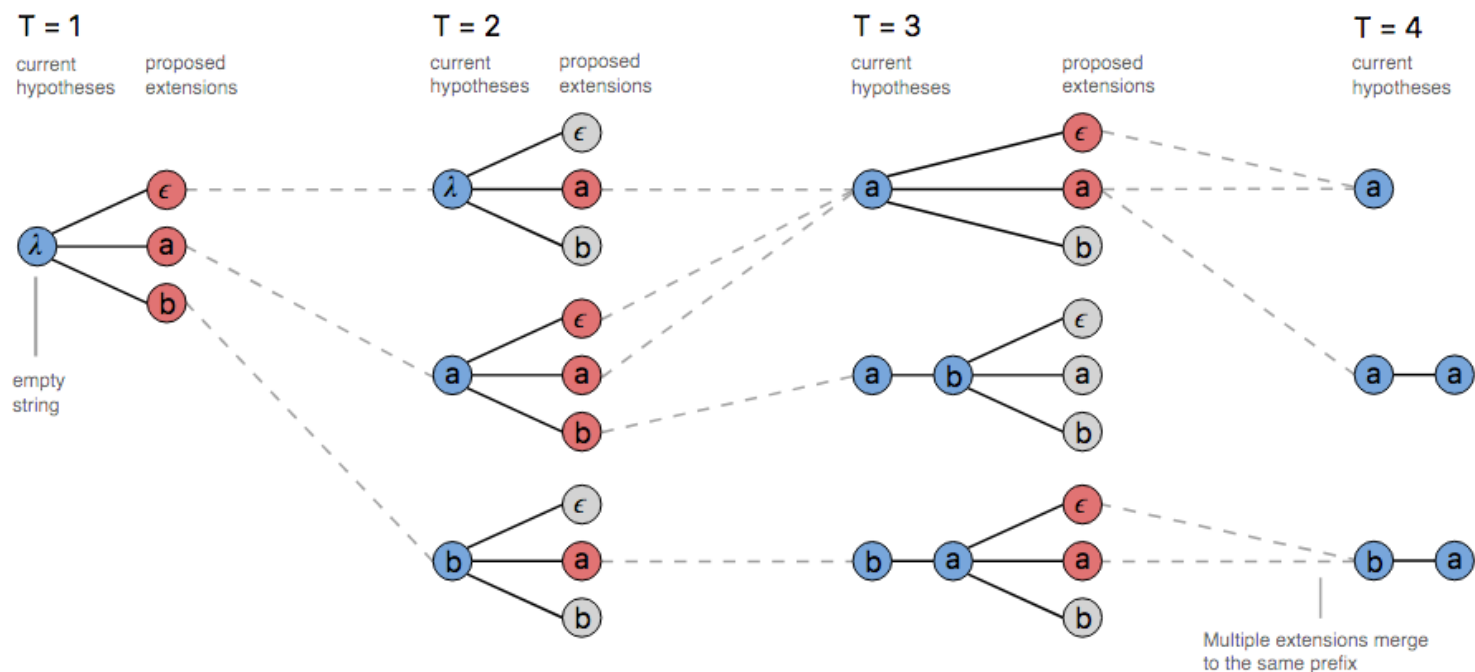
- not taking into account CTC function property:
many paths correspond to one labelling





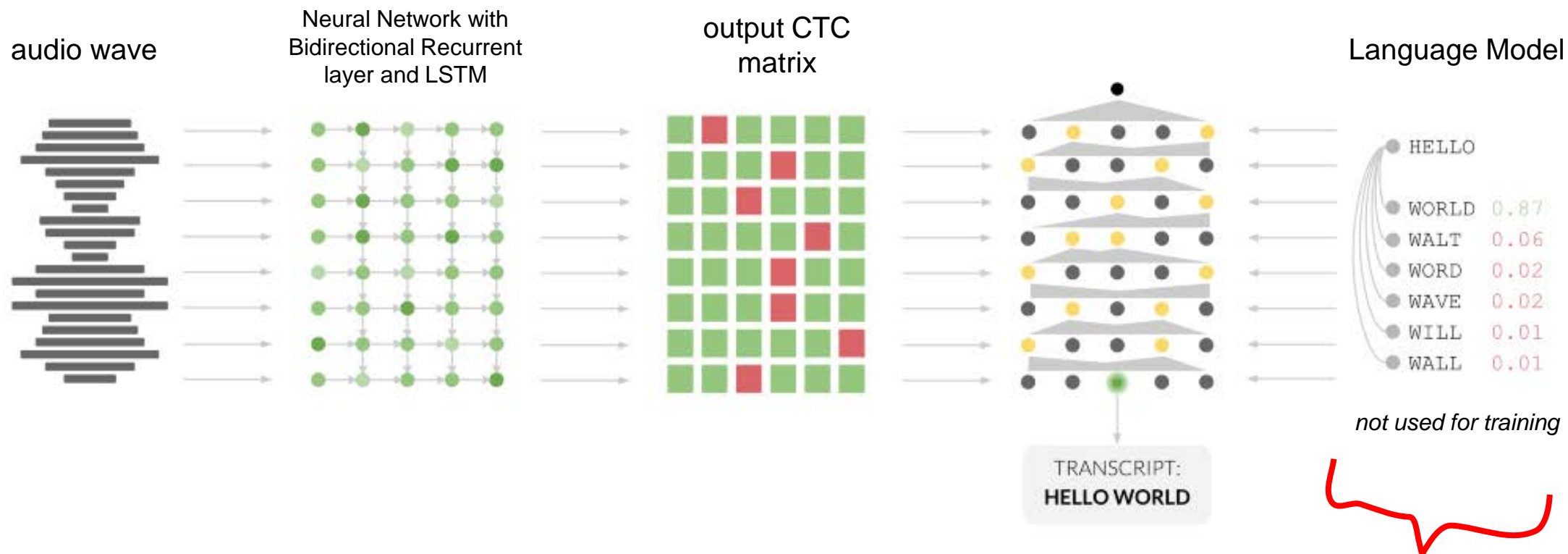
How to decode trained network output?

CTC Beam (Prefix) search



- exploring different paths and selecting at each time step N (N = beam_width) best (max probability)
- when reaching **space** symbol score sequence of words using Language Model

Language model





Language model estimation using KenLM

- language model is estimated from books and Wikipedia texts using KenLM toolkit
- consists of 1,2,3,4-grams
- pruned to 10 minimum n-gram usage

```

\data\
ngram 1=45752
ngram 2=204526
ngram 3=276301
ngram 4=254665

```

```

\1-grams:
-5.3364944 <unk> 0
0 <s> -0.5964499
-1.0091884 </s> 0
-4.122425 переходит -0.41616163
-1.804727 в -0.43273103
-3.5829756 корпус -0.32907426
-4.122425 выступает -0.21588245
-2.9297626 даже -0.27214578
-5.0543957 заместитель -0.079595566
-4.4398475 министра -0.26097438
-4.576069 финансов -0.079595566
-4.0951505 сергей -0.1420494
-5.1945915 шаталов -0.079595566
-1.6386111 и -0.3257205
-3.8569946 следующий -0.13813558
-5.1945915 мотив -0.079595566
-2.3505497 как -0.41405472
-3.2637522 бы -0.18714389
-4.775645 кровать -0.079595566
-3.2637522 сделать -0.25918472
-4.336306 получше -0.25343436
-2.614018 очень -0.35119596
-2.9444592 тоже -0.2585964
-4.576069 удобная -0.22002326
-3.7866316 вещь -0.30276018
-2.6036654 он -0.30190697
-4.576069 предназначен -0.21715578
-2.3278315 для -0.34708676
-1.9070399 на -0.3659052
-3.8134248 которой -0.16508213
-2.3524578 я -0.4461375
-5.1945915 высаживала -0.079595566
-4.336306 рассаду -0.20094381
-4.775645 комплексе -0.079595566

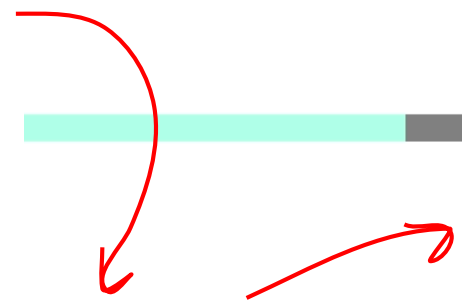
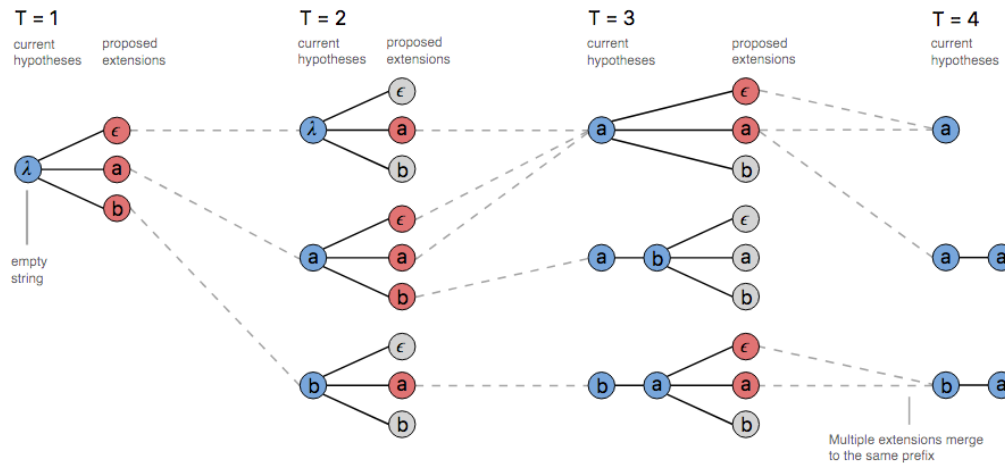
```

```

-1.1593988 объявлен в международный -0.017275982
-0.9171848 футболисте проводится международный -0.017275982
-1.1607316 атмосферу в женский -0.017275982
-3.1798248 так и женский -0.017275982
-1.461442 реагируют на женский -0.017275982
-3.4357328 что это женский -0.017275982
-1.2652726 в международный женский -0.017275982
-0.9989899 <s> носок женский -0.017275982
-2.308173 <s> поскольку оценивают -0.017275982
-1.1604284 риски как неприемлемые -0.017275982
-1.4380921 пока вода растворяет -0.017275982
-1.091364 <s> работала массивовано -0.017275982
-2.280585 сравнению с моделями -0.017275982
-1.8502711 сравнить с моделями -0.017275982
-0.96647924 всеми современными моделями -0.017275982
-1.1823485 более доступными моделями -0.017275982
-1.0221285 <s> новыми моделями -0.017275982
-0.8326845 между бюджетными моделями -0.017275982
-1.7620627 никакой не придуманный -0.017275982
-0.64793694 не придуманный кремлевскими -0.017275982
-0.64793694 придуманный кремлевскими троллями -0.017275982
-4.0215535 <s> на собранной -0.017275982
-1.4228556 и значительно сокращает -0.017275982
-1.1824195 быстрая стирка сокращает -0.017275982
-1.114462 сокращает объем получаемого -0.017275982
-1.1608133 получаемого и передаваемого -0.017275982
-1.1605312 сцены с русскими -0.017275982
-1.9169104 <s> написано русскими -0.017275982
-0.6479288 <s> печатаю русскими -0.017275982
-2.0575116 думаю вы ничем -0.017275982
-1.1602981 высокому все ничем -0.017275982
-1.1218724 уже ремонт балки -0.017275982
-1.1608133 плюсах и минусах -0.017275982
-1.1824267 <s> пещера вымытая -0.017275982
-1.1607751 вымытая в скале -0.017275982
-1.0221285 достаточно холодный морской -0.017275982
-0.6479141 в скале морской -0.017275982
-1.3439243 <s> местной морской -0.017275982
-0.6479141 рубежу батальона морской -0.017275982
-0.64793694 <s> оптовая биржевая -0.017275982

```

How language model is used?



-1.1593988	объявлен в международный	-0.017275982
-0.9171848	футболисте проводится международный	-0.017275982
-1.1607316	атмосферу в женский	-0.017275982
-3.1798248	так и женский	-0.017275982
-1.461442	реагируют на женский	-0.017275982
-3.4357328	что это женский	-0.017275982
-1.2652726	в международный женский	-0.017275982
-0.9989899	<s> носок женский	-0.017275982
-2.308173	<s> поскольку оценивают	-0.017275982
-1.1604284	риски как неприемлемые	-0.017275982
-1.4380921	пока вода растворяет	-0.017275982
-1.091364	<s> работала массивовано	-0.017275982
-2.280585	сравнению с моделями	-0.017275982
-1.8502711	сравнивать с моделями	-0.017275982
-0.96647924	всеми современными моделями	-0.017275982
-1.1823485	более доступными моделями	-0.017275982
-1.0221285	<s> новыми моделями	-0.017275982
-0.8326845	между бюджетными моделями	-0.017275982
-1.7620627	никакой не придуманный	-0.017275982
-0.64793694	не придуманный кремлевскими	-0.017275982
-0.64793694	придуманый кремлевскими троллями	-0.017275982
-4.0215535	<s> на собранной	-0.017275982
-1.4228556	и значительно сокращает	-0.017275982
-1.1824195	быстрая стирка сокращает	-0.017275982
-1.114462	сокращает объем получаемого	-0.017275982
-1.1608133	получаемого и передаваемого	-0.017275982
-1.1605312	сцены с русскими	-0.017275982
-1.9169104	<s> написано русскими	-0.017275982
-0.6479288	<s> печатаю русскими	-0.017275982
-2.0575116	думаю вы нипочем	-0.017275982
-1.1602981	высоцкому все нипочем	-0.017275982
-1.1218724	уже ремонт балки	-0.017275982
-1.1608133	плюсах и минусах	-0.017275982
-1.1824267	<s> пещера вымытая	-0.017275982
-1.1607751	вымытая в скале	-0.017275982
-1.0221285	достаточно холодный морской	-0.017275982
-0.6479141	в скале морской	-0.017275982
-1.3439243	<s> местной морской	-0.017275982
-0.6479141	рубжу батальона морской	-0.017275982
-0.64793694	<s> оптовая биржевая	-0.017275982
-1.0221285		

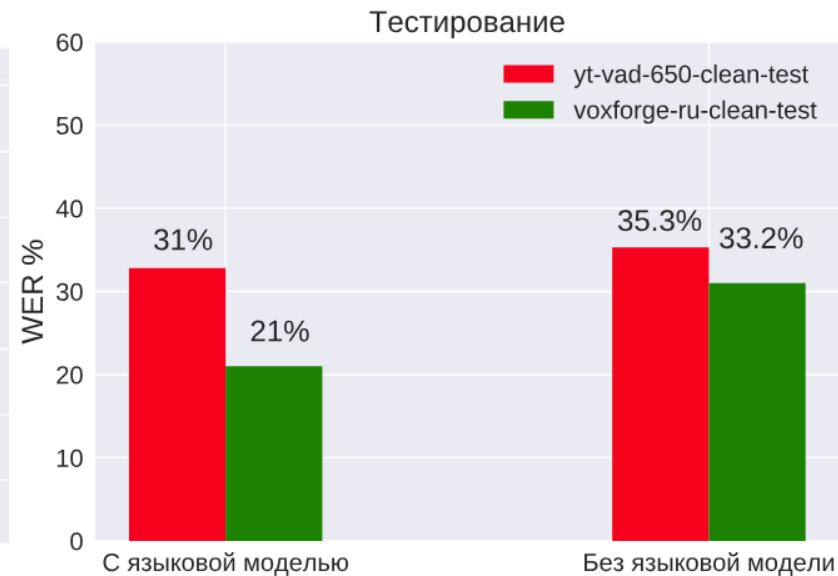
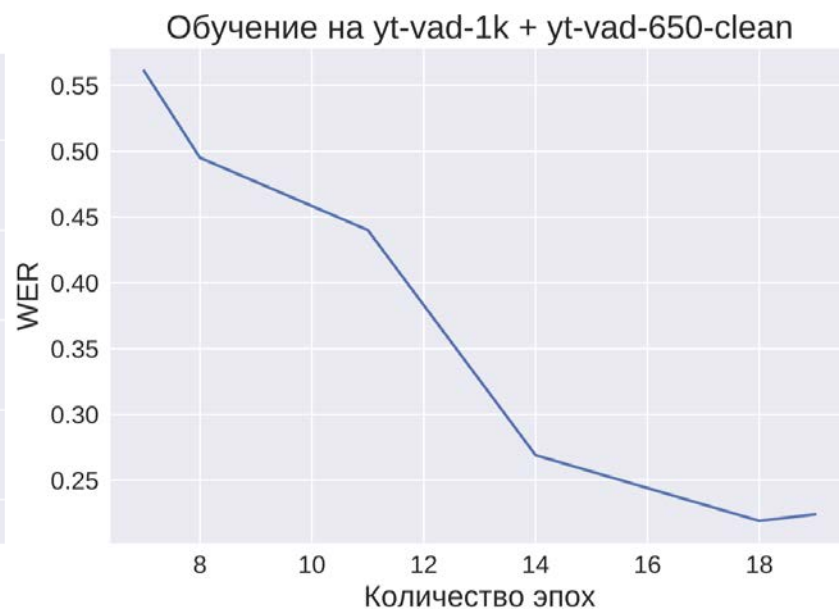
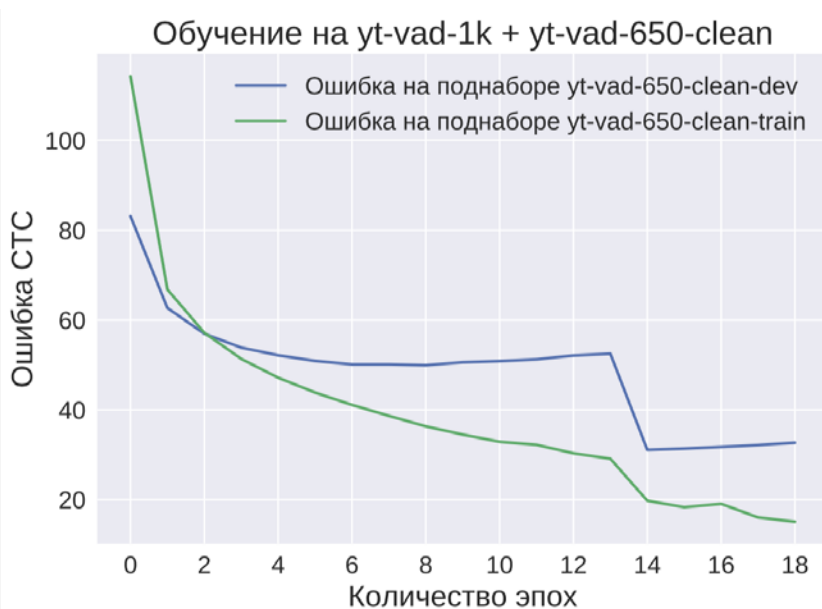
During beam search decoding, if next decoded symbol is **space**

- LM is queried for getting probability of currently decoded word sequence
- sequence with bigger probability gets more score during decoding

Training



yt-vad-1k (1000h) + yt-vad-650-clean



- Min WER **21%** on the **voxforge-ru-clean-test**

Search example



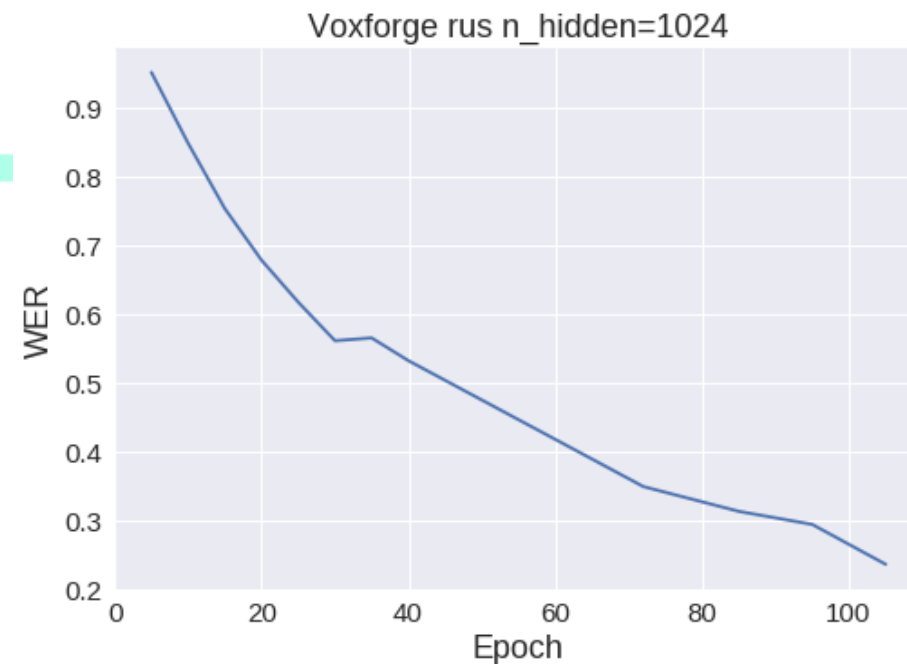
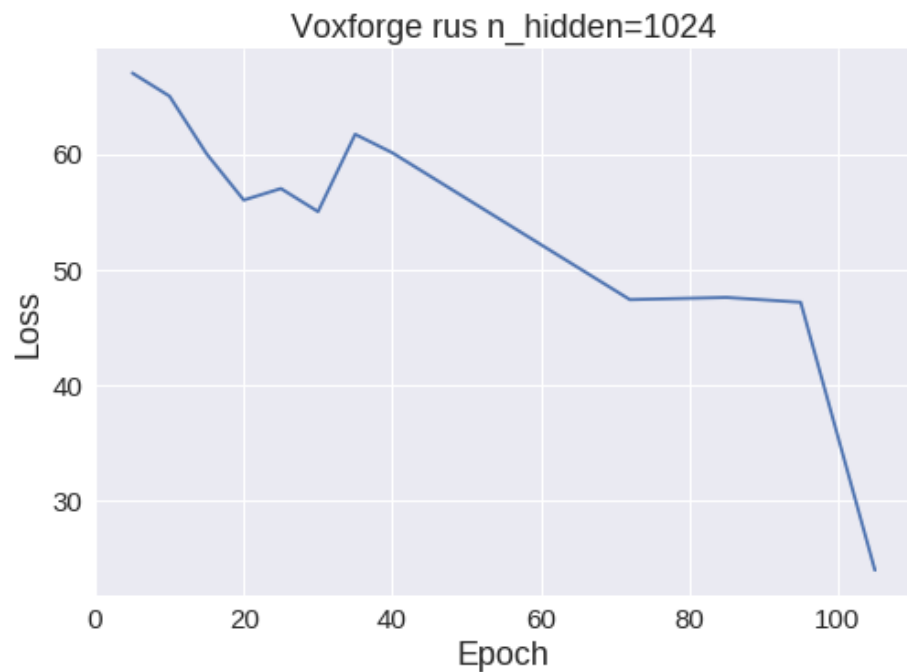
deepspeech search





Next researches

Training on small dataset VoxForge_ru ~ 26 hours



- Training time:** ~ 4 min per epoch (on 2 x Tesla P100)
- Testing time (CPU beam search):** ~ 15 min
- current beam search implementation runs on CPU due to querying of KenLM and lacks multithreading



Thank you for attention!