



Building corpora of transcribed speech from open access sources

George Fedoseev, Oleg Jakushkin,
Anna Shaleva















Collecting speech data

Collecting speech data



Getting pairs (audio, transcript) from open source dataset VoxForge.org



Name	Last modified	Size
 Parent Directory		-
 1-20121125-pgp.tgz	2014-05-28 13:49	2.1M
 1981-20120705-haq.tgz	2014-05-22 04:22	1.9M
 1981-20120705-rjp.tgz	2014-05-22 04:22	1.4M
 1981-20120706-azq.tgz	2014-05-22 04:22	1.6M
 1981-20120706-hpa.tgz	2014-05-22 04:22	1.8M
 1981-20120706-kwo.tgz	2014-05-22 04:22	1.7M
 1981-20120706-rxa.tgz	2014-05-22 04:22	2.1M
 1981-20120706-rya.tgz	2014-05-22 04:22	2.0M
 1981-20120706-vmc.tgz	2014-05-22 04:22	1.7M
 1981-20120706-zfp.tgz	2014-05-22 04:22	2.0M
 4ertus2-20101217-zoo.taz	2014-05-23 04:41	2.2M

Total Minutes	Total Hours
-----	-----
1490.8	24.8

Collecting speech data

Getting pairs (audio, transcript) from YouTube videos with captions



Collecting speech data



author-provided subtitles

```
00:00:48.110 --> 00:00:50.500
эта зависимость растёт
быстрее,

00:00:50.520 --> 00:00:53.980
где она убывает быстрее, где медленнее...

00:00:53.990 --> 00:00:55.210
Если мы возьмём
```

validate author subtitles
using auto-generated
YouTube subtitles
+ more precise audio cutting
using timing from auto-subs

start and end time for
each word in video

+ autogenerated YouTube subtitles

```
00:00:48.989 --> 00:00:50.299 align:start position:0%
<c.colorCCCCCC>растет</c><c.colorE5E5E5><00:00:49.440><c> быстрее</c><00:00:49.980><c> где</c></c>

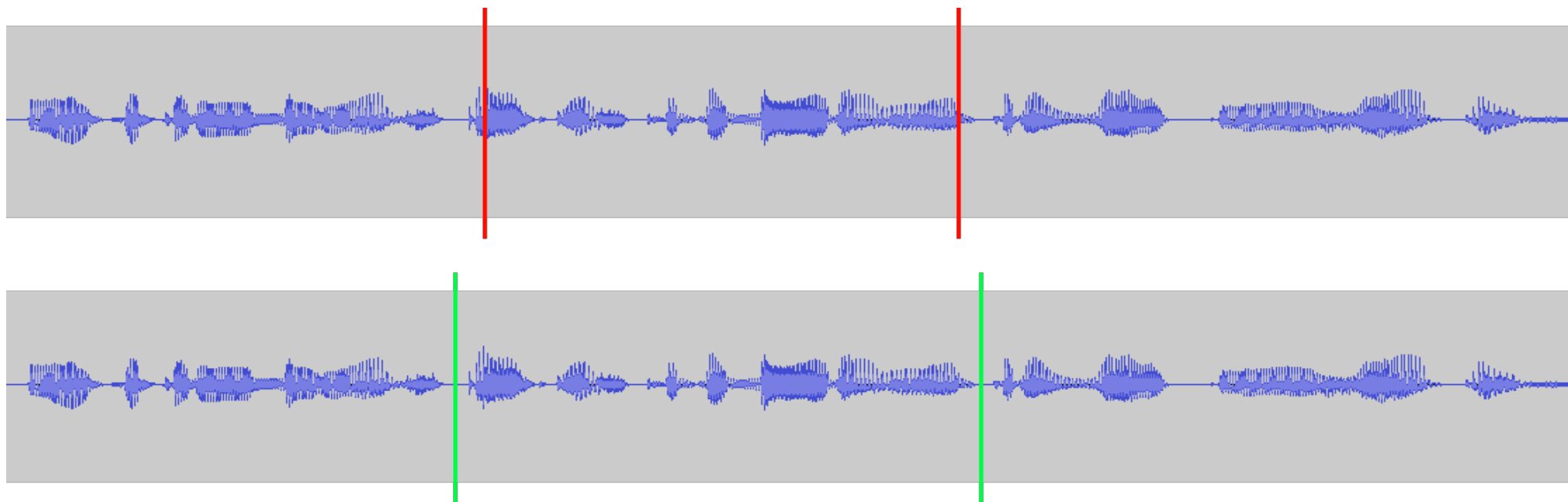
00:00:50.309 --> 00:00:51.139 align:start position:0%
она<00:00:50.489><c> убывает</c><00:00:50.730><c> быстрее</c></c>

00:00:51.149 --> 00:00:54.229 align:start position:0%
где<c.colorE5E5E5><00:00:51.809><c> медленнее</c></c><c.colorCCCCCC><00:00:53.239><c> все</c></c>
```

Collecting speech data



Correction of words cutting using WebRTC Voice Activity Detector



<https://github.com/wiseman/py-webrtcvad>

Collecting speech data

Getting pairs (audio, transcript) from Echo of Moscow radio shows

The screenshot displays four interview cards from the Echo of Moscow website. Each card includes a guest photo, name, title, a transcript snippet, and an audio player interface with download options.

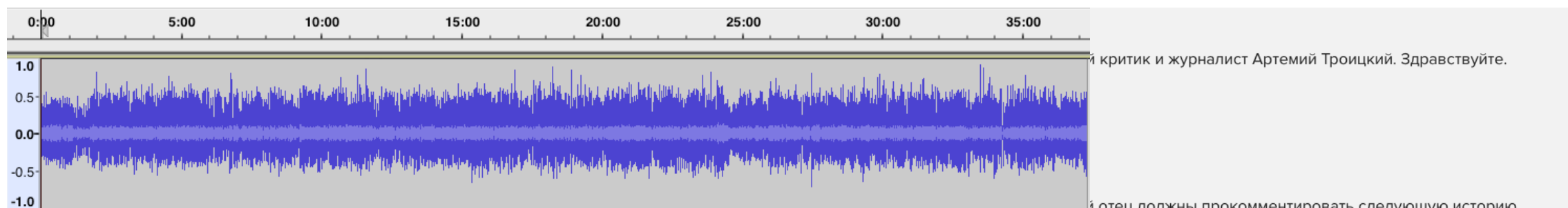
- Card 1:** Interview with **Константин Ремчуков**, главный редактор «Независимой газеты». Transcript snippet: "У латышей и эстонцев аргументы понятные. Если вы живете в стране, вы, естественно, должны знать госязык. Иначе вы будете не совсем полноценными гражданами. Вопрос: где и как его учить. Но ответа практически нет...". Audio player: 29 comments, 6748 views, 19:08 duration, 39:08 listening time, 9.0 MB download.
- Card 2:** Interview with **Артемий Троицкий**, журналист. Transcript snippet: "У латышей и эстонцев аргументы понятные. Если вы живете в стране, вы, естественно, должны знать госязык. Иначе вы будете не совсем полноценными гражданами. Вопрос: где и как его учить. Но ответа практически нет...". Audio player: 47 comments, 13118 views, 17:08 duration.
- Card 3:** Interview with **Владимир Гельман**, политолог. Title: "Особое мнение СПб". Audio player: 0 comments, 109 views, 11:11 duration.
- Card 4:** Interview with **Максим Шевченко**, журналист. Transcript snippet: "Путин в Кемерове был растерян. Я впервые видел президента, который запинается. Он там говорил, и было видно, что у него от внутреннего потрясения, по крайней...". Audio player: 61 comments, 49119 views, 30:03 duration.

Additional details from the first card: "ИНТЕРВЬЮ / Особое мнение", "гость: Константин Ремчуков", "политика, власть, государство, книги, видео", "Им в принципе".

Collecting speech data



Getting pairs (audio, transcript) from Echo of Moscow radio shows



- each episode has audio and transcript
- audio and text are **not aligned** as it is with YouTube subtitles
- audio has *intros* and *ads* in it which are not transcribed

критик и журналист Артемий Троицкий. Здравствуйте.

отец должны прокомментировать следующую историю.

Президент Латвии подписал поправки в законодательство, предусматривающие постепенный перевод школьного обучения на латышский язык. Соответствующие документы опубликованы в понедельник в официальном издании страны. Это та самая история с русскоязычными школами, которая закончится довольно быстро.

А. Троицкий

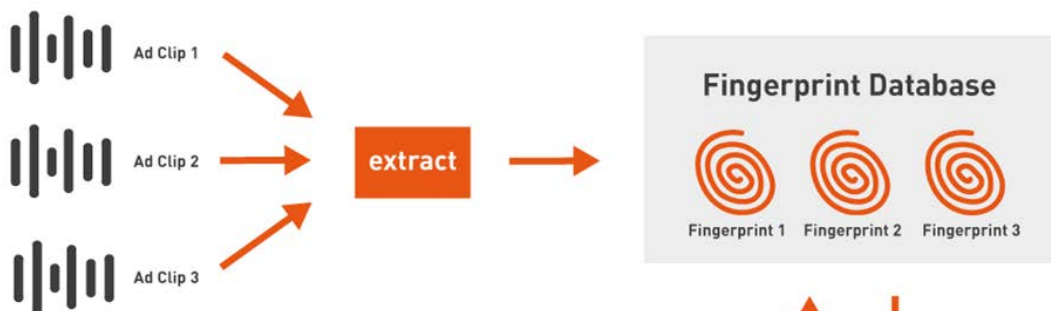
- Да, это очень тяжелая тема, которая дискутируется уже очень много лет. Аргументы есть у всех сторон. У латышей и эстонцев аргументы, в общем, очень понятные. То есть если вы живете в стране Латвии или Эстонии или Литве или Белоруссии и так далее, то вы естественно, должны знать государственный язык этой страны. Иначе вы будете не совсем полноценными гражданами. Для того чтобы вы знали этот язык этой страны, естественно, вам нужно его учить. Далее встает вопрос: где и как его учить. И на этот вопрос ответа практически нет. Потому что сейчас я хуже знаю ситуацию в Латвии, разумеется. В Эстонии я знаю очень хорошо. В Эстонии имеется большое количество русских школ, и в Таллине я уже не говорю о Нарве преподавание эстонского языка там поставлено

Collecting speech data

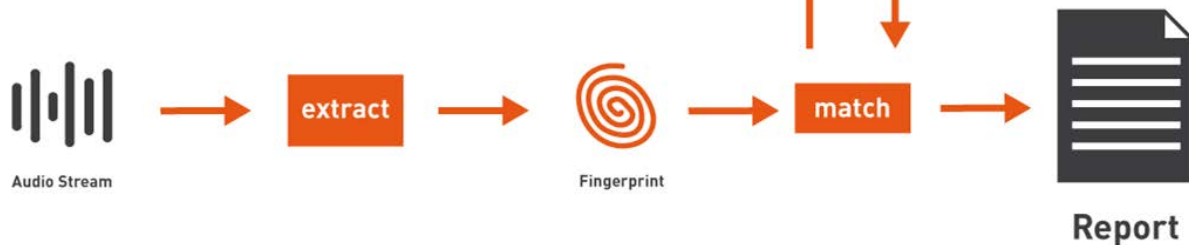


Removing intros from audio

Fingerprint Database Creation



Content Identification



<https://github.com/dpwe/audfprint>

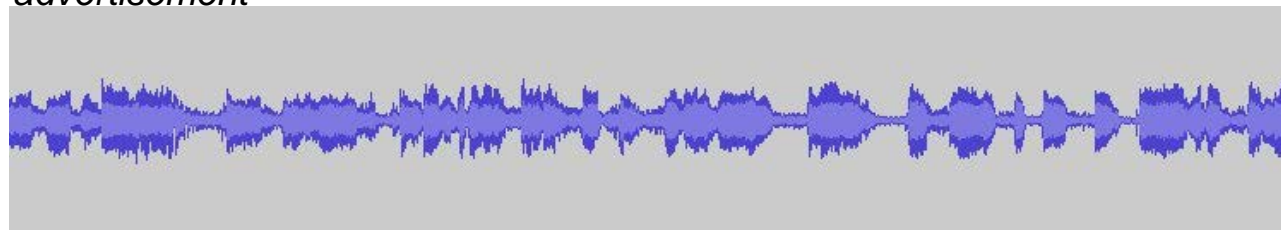
Collecting speech data



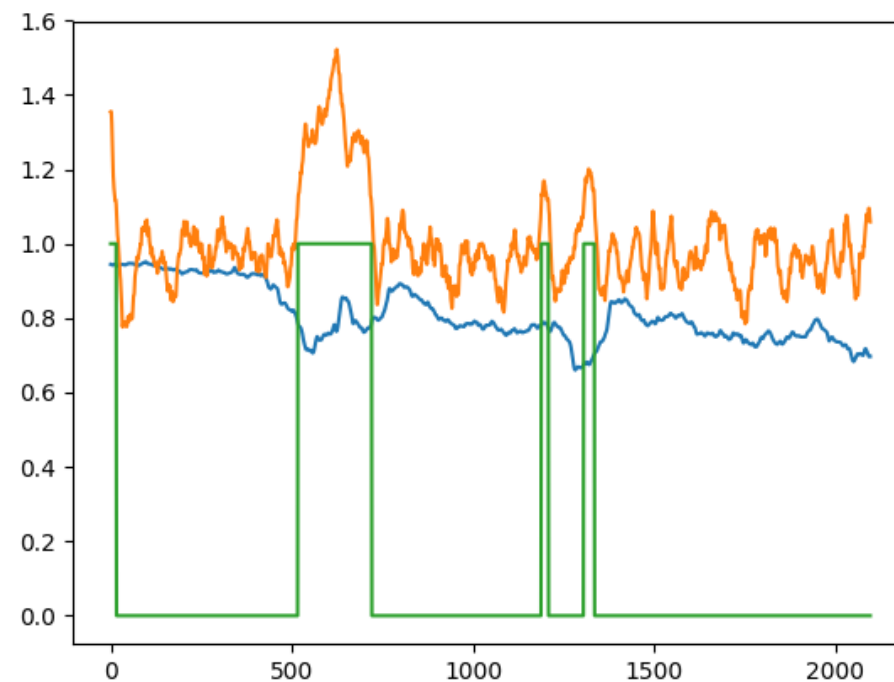
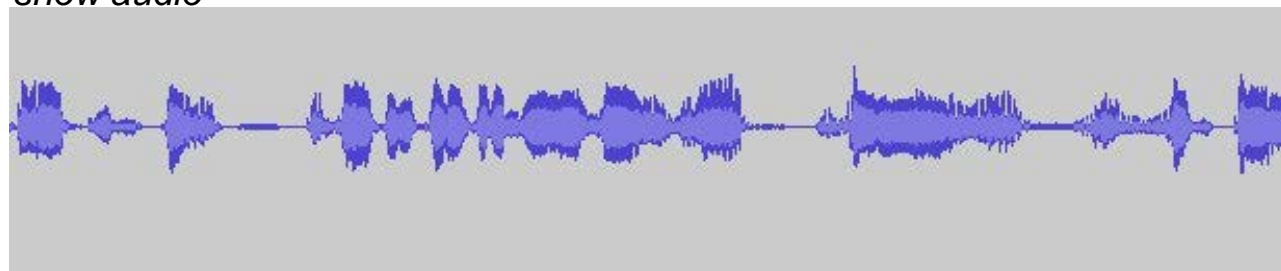
Removing ads and intros from audio

Volume and density of speech ads detection

advertisement



show audio

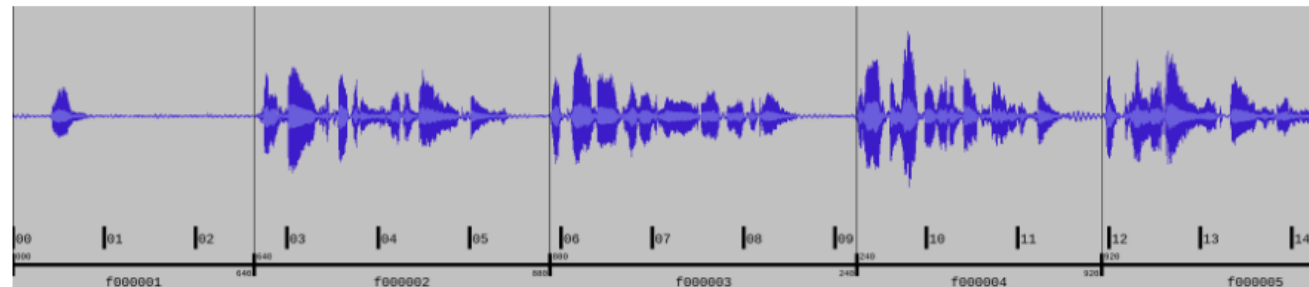




Collecting speech data

Forced alignment: text -> no_ads_audio

1	=> [00:00:00.000, 00:00:02.640]
From fairest creatures we desire increase,	=> [00:00:02.640, 00:00:05.880]
That thereby beauty's rose might never die,	=> [00:00:05.880, 00:00:09.240]
But as the ripper should by time decease,	=> [00:00:09.240, 00:00:11.920]
His tender heir might bear his memory:	=> [00:00:11.920, 00:00:15.280]
But thou contracted to thine own bright eyes,	=> [00:00:15.280, 00:00:18.800]
Feed'st thy light's flame with self-substantial fuel,	=> [00:00:18.800, 00:00:22.760]
Making a famine where abundance lies,	=> [00:00:22.760, 00:00:25.680]
Thy self thy foe, to thy sweet self too cruel:	=> [00:00:25.680, 00:00:31.240]
Thou that art now the world's fresh ornament,	=> [00:00:31.240, 00:00:34.400]
And only herald to the gaudy spring,	=> [00:00:34.400, 00:00:36.920]
Within thine own bud buriest thy content,	=> [00:00:36.920, 00:00:40.640]
And tender churl mak'st waste in niggarding:	=> [00:00:40.640, 00:00:43.640]
Pity the world, or else this glutton be,	=> [00:00:43.640, 00:00:48.080]
To eat the world's due, by the grave and thee.	=> [00:00:48.080, 00:00:53.240]



<https://github.com/readbeyond/aeneas/>

Collecting speech data



Problems with speech data from radio shows

- repeated words in conversation, which are not in transcript
- thinking sounds “mmm, hmm, eem”, which are not in transcript
- speakers interrupt each other



Due to above reasons automatic forced alignment algorithm produces alignments with error of 1-3 words, making this dataset not useable for training for now.

Possible solution would involve using speech recognizer to validate slicing of forced alignment algorithm.



High Performance Computing Node

CPU: 2 x Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz

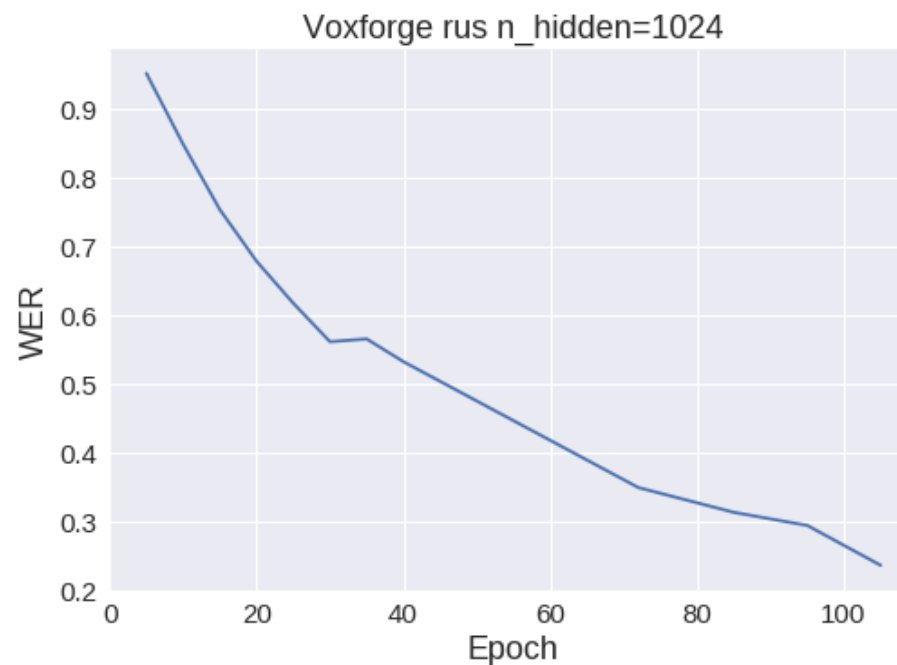
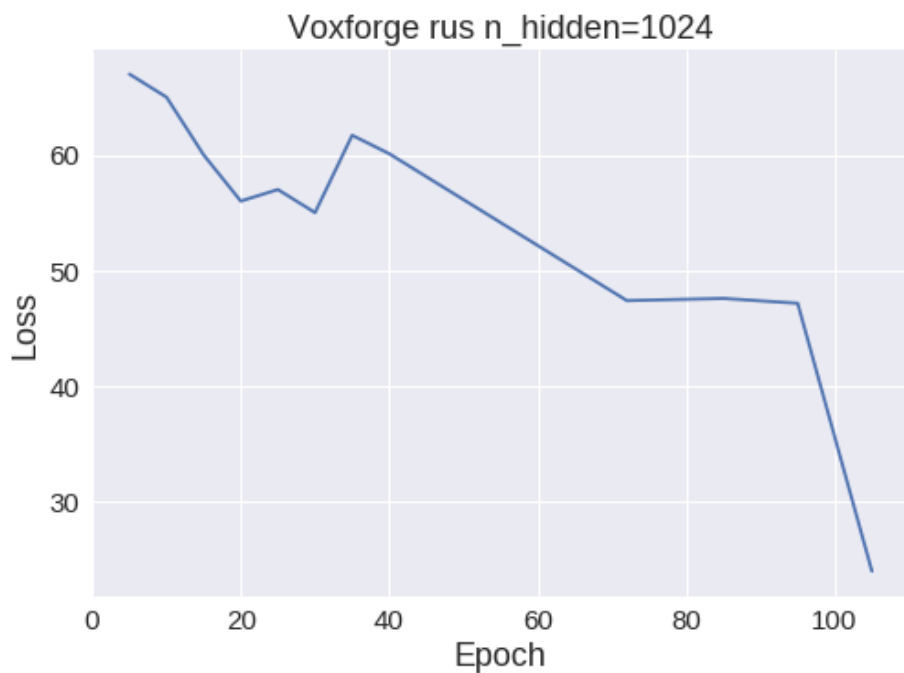
RAM: 256 GB

GPU: 2 x Nvidia Tesla P100 16GB

Training



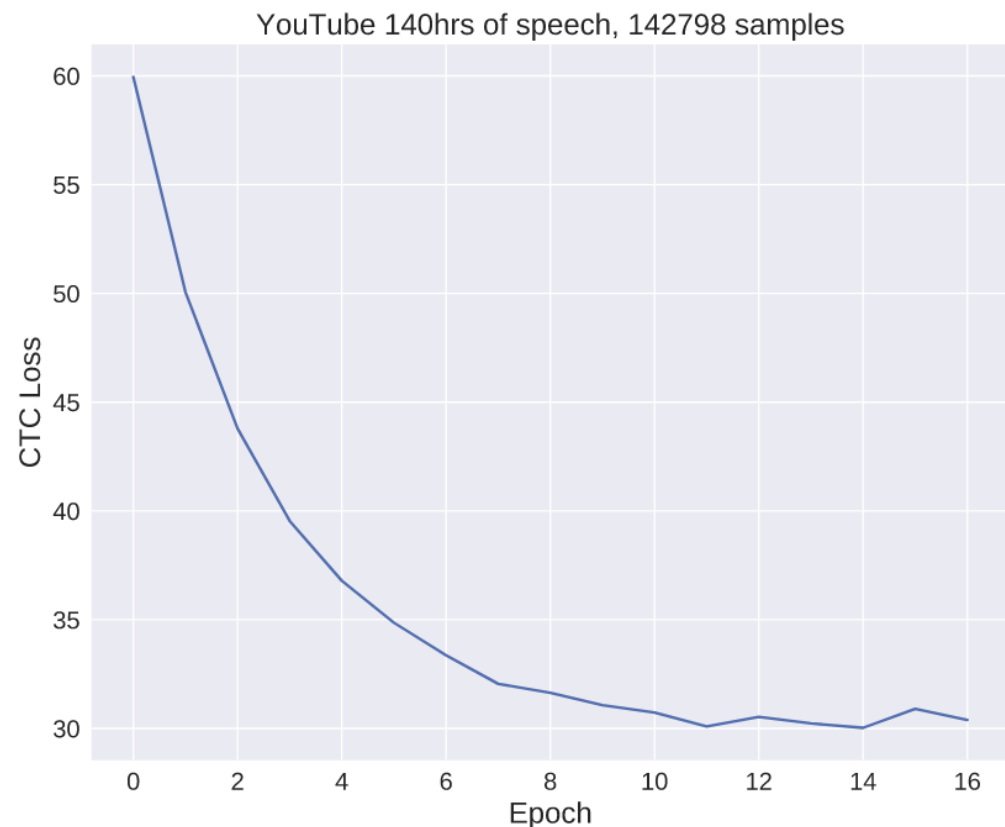
Training on small dataset VoxForge_ru ~ 26 hours



- 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



Training on YouTube captions dataset ~ 140 hours



Training time: ~ 20m per epoch
on 2 x Tesla P100

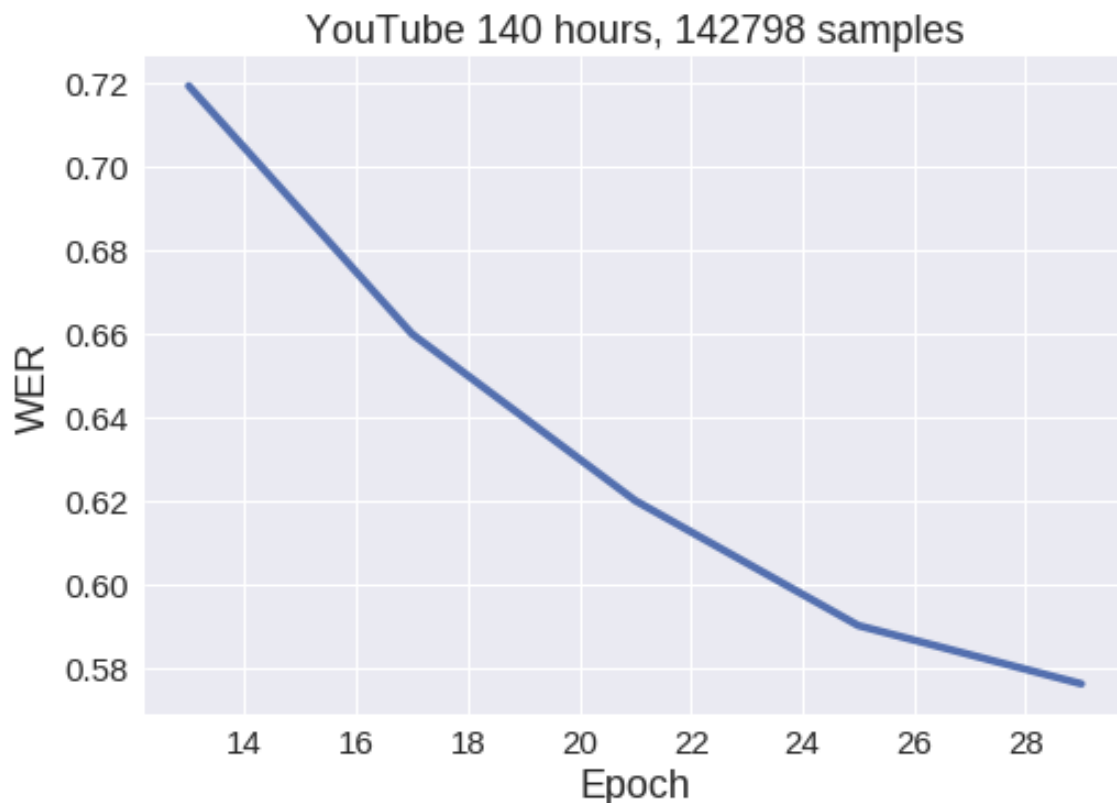
Testing time (CPU beam search):
1.5h
(beam_width=1024)

thousands of speakers, **noisy** recording environment - **harder to converge** for NN

Training



Training on YouTube captions dataset ~ 140 hours



less = better

work in progress
to reduce more

thousands of speakers, **noisy** recording environment - **harder to converge** for NN

Speech recognition examples



YouTube captions dataset ~ 140 hours

```
I -----  
I WER: 0.100000, loss: 2.050212, mean edit distance: 0.040816  
I - src: "он дает гарантии на эти контракты на любые из них"  
I - res: "он дает гарантии на эти контакты на любые из них "  
I -----  
I WER: 0.111111, loss: 0.668558, mean edit distance: 0.052632  
I - src: "давайте посмотрим на исходную картинку и на наш результат"  
I - res: "давайте посмотрим на исходную картинку и на наш и результат "  
I -----  
I WER: 0.111111, loss: 0.927856, mean edit distance: 0.038462  
I - src: "личную точку зрения то есть не отражает сути явления"  
I - res: "личную точку зрения то есть не отражает суть явления "  
I -----  
I WER: 0.111111, loss: 1.971290, mean edit distance: 0.050000  
I - src: "до проблема именно в них это один из ста"  
I - res: "до проблемы именно в них это один из ста "  
I -----  
I WER: 0.111111, loss: 2.242571, mean edit distance: 0.051282  
I - src: "от того что вы не тратите силы на семью"  
I - res: "от того что вы не тратить силы на семью "  
I -----  
I WER: 0.111111, loss: 2.246983, mean edit distance: 0.043478  
I - src: "вот здесь мы с вами и будем поднимать резкость"  
I - res: "вот здесь мы с вами и будем понимать резкость "  
I -----
```

average WER 58%



Thank you for attention!