НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Speech synthesis on a shoe string

Francis M. Tyers†

ftyers@hse.ru
https://www.hse.ru/org/persons/209454856

Национальный исследовательский университет
«Высшая школа экономики» (Москва)

† with Josh Meyer, Jonathan North Washington and Marinyé Andreeva

Introductions **Motivation**

Project **Methdology**

Introductions

6 An endangered language will progress if its speakers can make use of electronic technology

Main research interest:

- Language technology for maintenance and revivial

I have worked with:

- Machine translation [Assimilation, Dissemination]
- Modelling morphology [Spellchecking]
- Syntactically-annotated corpora [Involve MLers]

And now:

- Speech synthesis

Motivation

Why speech synthesis?

- Interesting results with little data
- Lay groundwork for further speech-related work
- Real «application»

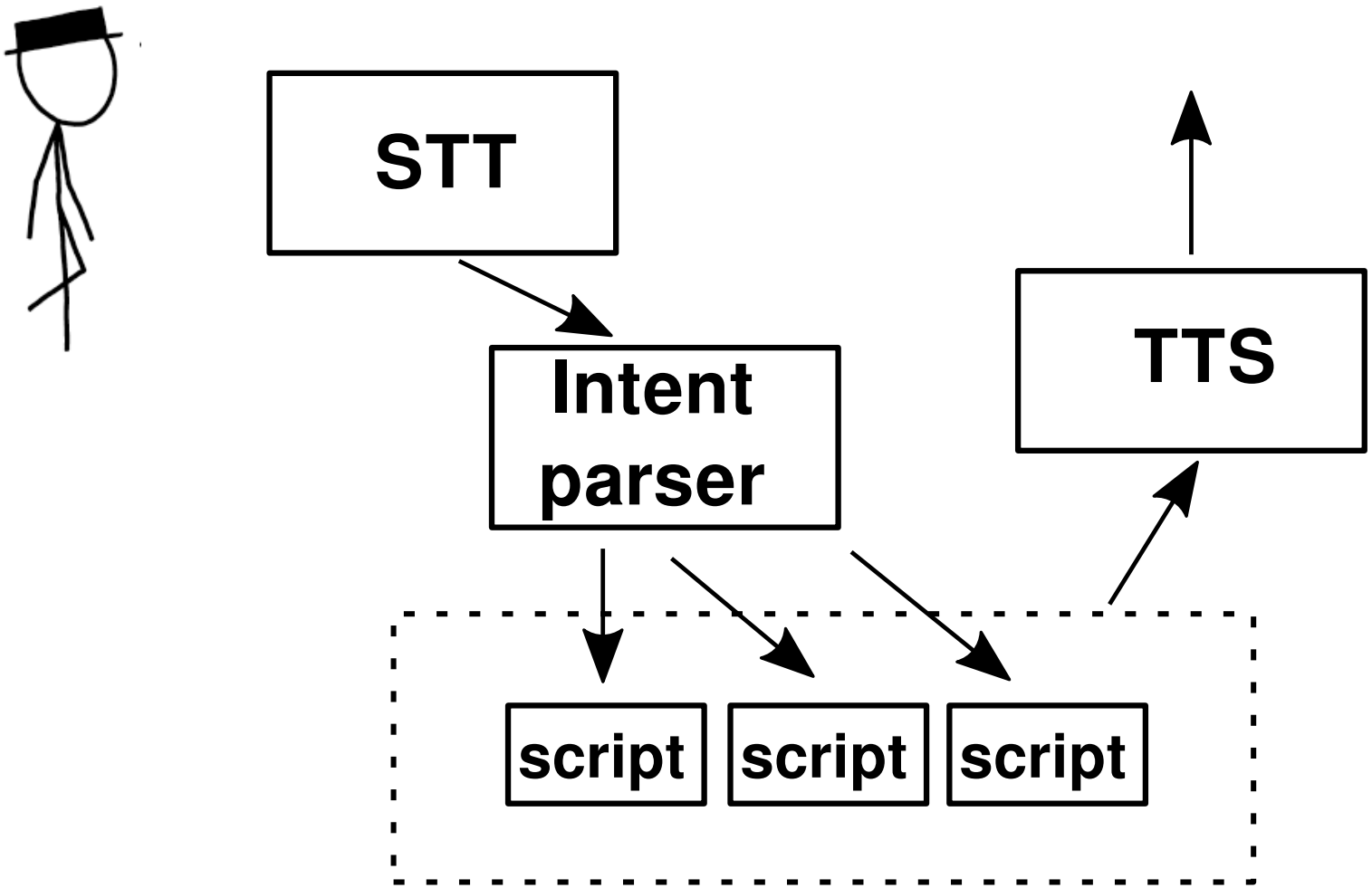
Data requirements

Synthesis (TTS)	**Recognition** (STT)
– Hours of audio	– 100s of hours of audio
– Studio quality	– Telephone quality
– Single speaker	– Many speakers
– Not processor intensive	– Processor intensive

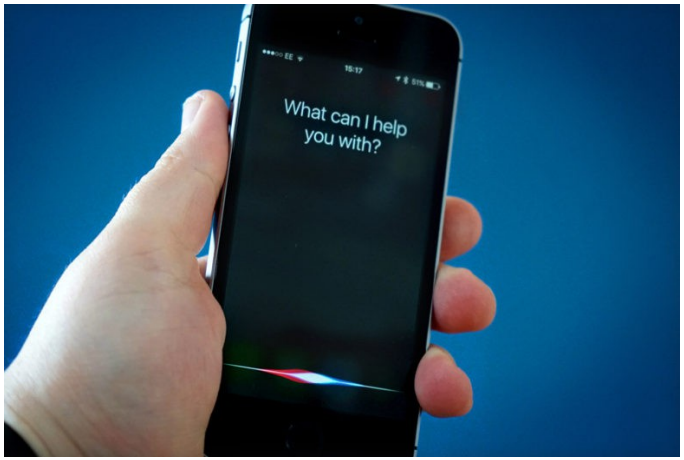
Virtual assistants

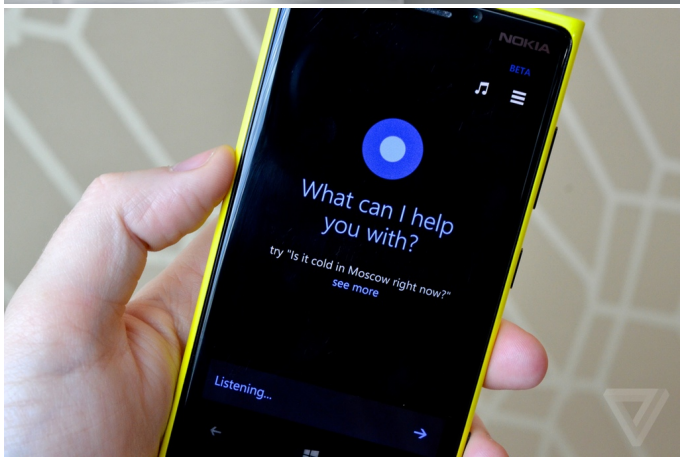
- Talk to your computer
  - Information retrieval
  - «Smart» home stuff
- 3x faster than typing
- People really use them!
  - …in English (or Russian)

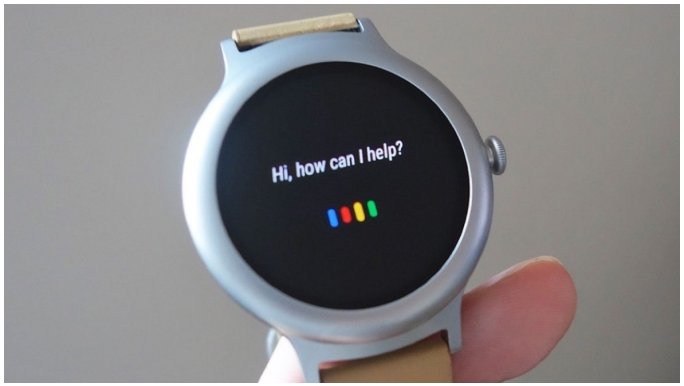
How they work



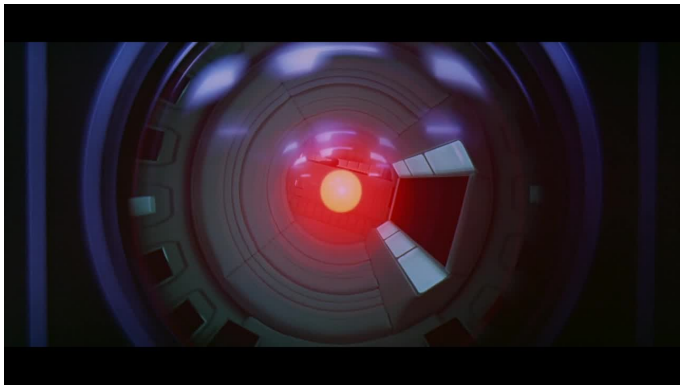

alexa:
arm away mode.
is the back door open?
what were the last 3 events?
just ask.
amazon echo x scout


What can I help you with?


NOKIA
BETA
What can I help you with?
try "Is it cold in Moscow right now?"
see more
Listening...


Hi, how can I help?


mycroft
An Artificial Intelligence for Everyone



Project description

Project

The idea:

- Collect a speech corpus of Chuvash
  - At least enough data to build a prototype
- Train a speech synthesis system
  - *Let's get the computer to speak Chuvash!*
- Develop an end-to-end "recipe"

Budget:

- Recordings: 7000 RUB (approx. 100€)
  - 10€/hour
  - Estimated 1 min of audio = 2 mins recording
  - 10 minutes/day:
    - 5 hours audio (10h recording) in 1 month
- Train ticket (MSK $\rightarrow$ SpB): 3000 RUB (approx. 50€)

Let's make a Chuvash voice!

Speech Synthesis Workshop, Higher School of Economics

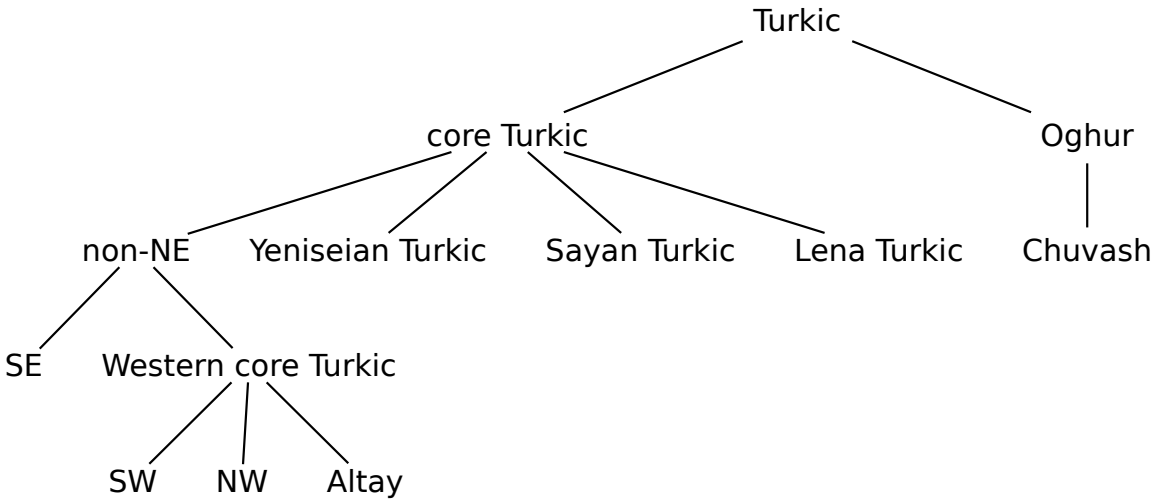
9th December, 2017

50€ well spent!

Participants

Follow Josh on Twitter! `@joshmeyerphd`

Chuvash

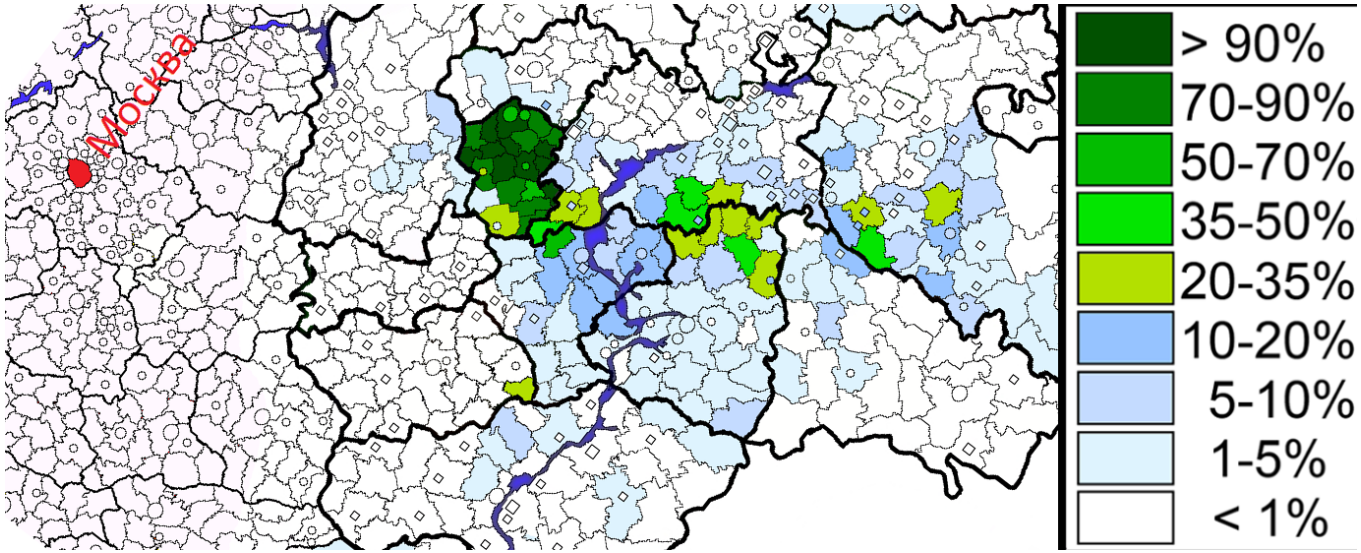


- Turkic language, Oghur sub-group
- Spoken in Chuvashia, Volga region, Russia
- Approximately one million speakers (c. 2010)
- Intergenerational transmission breaking down

Chuvash

- Turkic language, Oghur sub-group
- Spoken in Chuvashia, Volga region, Russia
- Approximately one million speakers (c. 2010)
- Intergenerational transmission breaking down

Chuvash



- Turkic language, Oghur sub-group
- Spoken in Chuvashia, Volga region, Russia
- Approximately one million speakers (c. 2010)
- Intergenerational transmission breaking down

Chuvash

- Turkic language, Oghur sub-group
- Spoken in Chuvashia, Volga region, Russia
- Approximately one million speakers (c. 2010)
- Intergenerational transmission breaking down

«Most respondents, even Chuvash-speakers, think that Chuvash is not modern.»[1]

- Sociolinguistic study of pupils and teachers
- First in Chuvashia

[1] Alòs i Font, H. (2016) “The Chuvash language in the Chuvash Republic: An example of the rapid decline of one of Russia’s major languages”

Methdology

Collecting data

chuvash.org:

- Online community
- Daily news in Chuvash
- Short articles 90–150 toks.
- CC-BY-SA licence

collection

- `wget`
- BeautifulSoup

Collecting data

chuvash.org:

- Online community
- Daily news in Chuvash
- Short articles 90–150 toks.
- CC-BY-SA licence

collection

- `wget`
- BeautifulSoup

Processing

Вика Сорокина 08.11.2017 12:13 | 244 хут пĕхнĕ

ПĂТĂРМАХСЕМ

Паян ирпе, çынсем ĕçе васканă самантра, Шупашкарти Мускав кĕперĕ çинче инкек сикссе тухнă. Çавна пула транспорт çӳресси çур сехетлĕхе чарăннине пĕлтереççĕ. 23 çулти çамрăк водитель тытса пыракан автобус сакăр машина çине пырса кĕнĕ.

Чăваш Енри Шалти ĕçсен министерствин сайтĕнче лекнĕ автобус 12-мĕш маршрутпа çӳренине асăннă. Транспортра 5 çын ларса пынă, анчах вĕсенчен нихăшĕ те аманманнине ĕнентереççĕ. Малтанласа палăртнă тăрăх, автобусăн тормозĕ ванма пултарнă. Анчах ку версие хаплĕхе çирĕплетме иртерех. Куракансем каланă тата водитель пĕлтернĕ тăрăх çапла пĕтĕмлетме май пур.

Инкек сăлтавне специалистсем палăртассипе çине тăраççĕ. Сă ӳкерчĕксем тăрăх хакласан, аваринче çăмăл машинăсем хытах лапчăнни курăнать.

#транспорт, #Шупашкар, #инкексем, #ШĔМ, #пăтăрмахсем, #çул-йĕр, #автобуссем, #Мускав кĕперĕ

Хыпар çăлăкçĕ: https://21.мвд.рф/news/item/11525823/

→

```
1 0001 ⟶ Паян ирпе, çынсем ĕçе
  васканă самантра, Шупашкарти
  Мускав кĕперĕ çинче инкек сиксе
  тухнă. ⏎
2 0002 ⟶ Çавна пула транспорт
  çӳресси çур сехетлĕхе чарăннине
  пĕлтереççĕ. ⏎
3 0003 ⟶ 23 çулти çамрăк водитель
  тытса пыракан автобус сакăр
  машина çине пырса кĕнĕ. ⏎
4 0004 ⟶ Чăваш Енри Шалти ĕçсен
  министерствин сайтĕнче лекнĕ
  автобус 12-мĕш маршрутпа
  çӳренине асăннă. ⏎
5 0005 ⟶ Транспортра 5 çын ларса
  пынă, анчах вĕсенчен нихăшĕ те
  аманманнине ĕнентереççĕ. ⏎
```

- Bespoke sentence segmenter in Python
- Fixing minor encoding errors (Latin vs. Cyrillic)

Audio recording materials

iPhone 4 S

- Recording device
- Noise insulation

Jonathan: « It’s amazing that the hardest part is managing to get the cushions to stand up straight »

Software

Recording:

- Auphonic (for iPhone 4S)

Processing:

- Google sheets
- `sox` – `http://sox.sourceforge.net/`
- pyAudioAnalysis – `https://github.com/tyiannak/pyAudioAnalysis`

Modelling:

- Ossian – `https://github.com/CSTR-Edinburgh/Ossian`
- Merlin – `https://github.com/CSTR-Edinburgh/merlin`

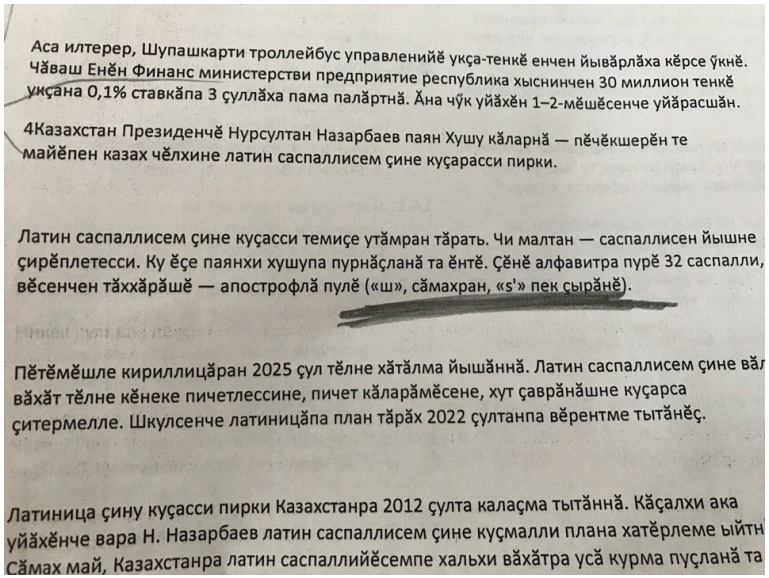
Recording

Our instructions:

- 5 seconds of silence at the beginning
  - For the noise profile
- 2 seconds between each sentence
  - To facilitate splitting

What Marinyé did:

- Produce a transcript
- Practice three times
- Read out numerals and abbreviations
- Correct errors in text — and note them down



Аса илтерер, Шупашкарти троллейбус управленийĕ укçа-тенкĕ енчен йывăрлăха кĕрсе ӳкнĕ. Чăваш Енĕн Финанс министерстви предприятие республика хыснинчен 30 миллион тенкĕ укçана 0,1% ставкăпа 3 çуллăха пама палăртнă. Ăна чӳк уйăхĕн 1–2-мĕшĕсенче уйăрасшăн.

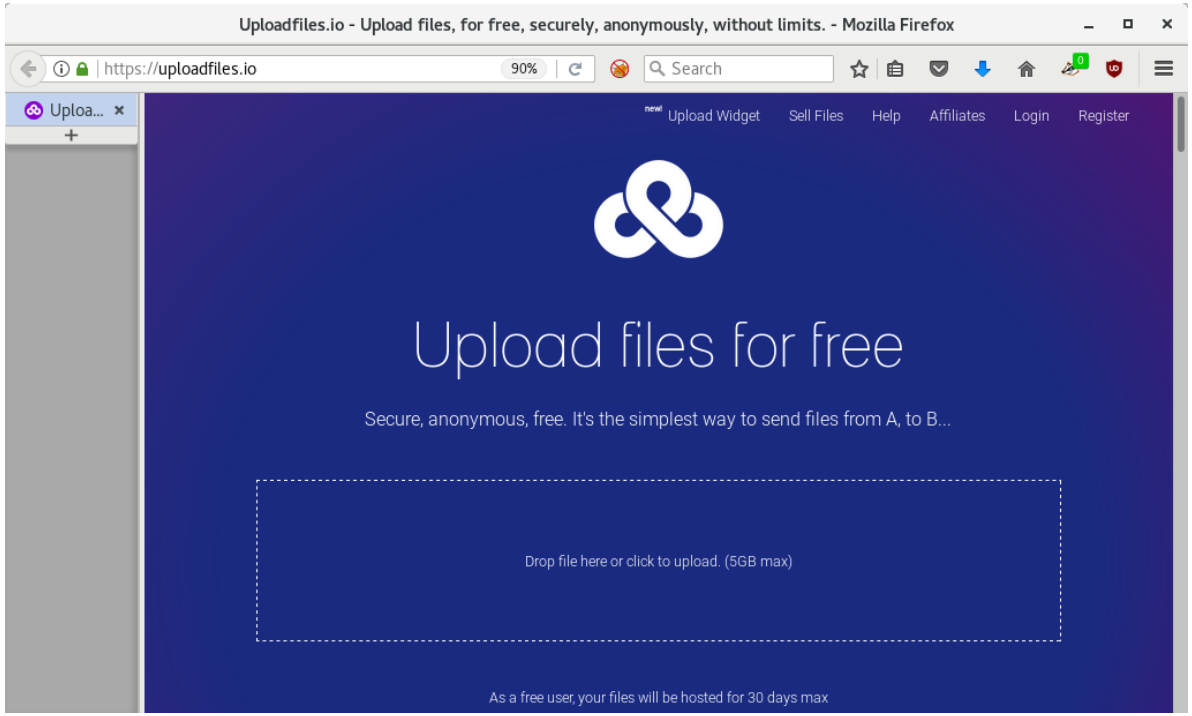
4Казахстан Президенчĕ Нурсултан Назарбаев паян Хушу кăларнă — пĕчĕкшерĕн те майĕпен казах чĕлхине латин саспаллисем çине куçарасси пирки.

Латин саспаллисем çине куçасси темиçе утăмран тăрать. Чи малтан — саспаллисен йышне çирĕплетесси. Ку ĕçе паянхи хушупа пурнăçланă та ĕнтĕ. Çĕнĕ алфавитра пурĕ 32 саспалли, вĕсенчен тăххăрăшĕ — апострофлă пулĕ («ш», сăмахран, «s'» пек сыпăнĕ).

Пĕтĕмĕшле кириллицăран 2025 çул тĕлне хăтăлма йышăннă. Латин саспаллисем çине вăл вăхăт тĕлне кĕнеке пичетлессине, пичет кăларăмĕсене, хут çаврăнăшне куçарса çитермелле. Шкулсенче латиницăпа план тăрăх 2022 çултанпа вĕрентме тытăнĕç.

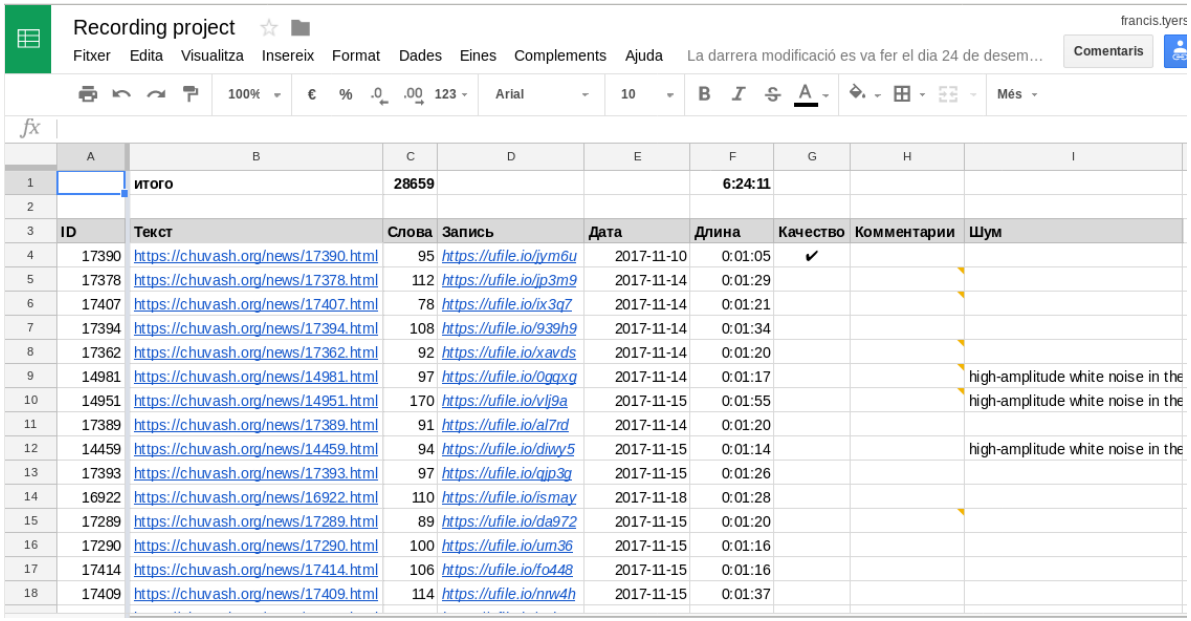
Латиница çину куçасси пирки Казахстанра 2012 çулта калаçма тытăннă. Кăçалхи ака уйăхĕнче вара Н. Назарбаев латин саспаллисем çине куçмалли плана хатĕрлеме ыйтн Сăмах май, Казахстанра латин саспаллийĕсемпе хальхи вăхăтра усă курма пуçланă та

Data storage and organisation



- `uploadfiles.io`
- Files last 30 days
- Easy to use

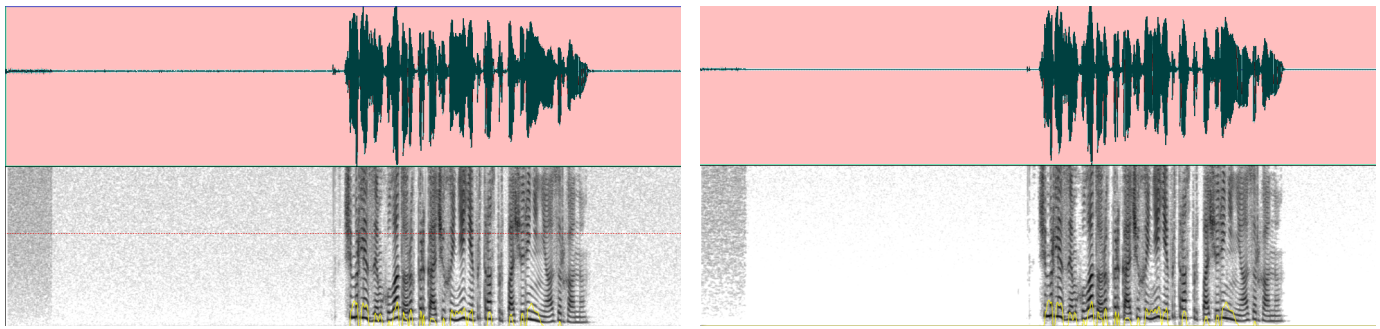
Data storage and organisation



	A	B	C	D	E	F	G	H	I
1		**итого**	**28659**			**6:24:11**			
2									
3	**ID**	**Текст**	**Слова**	**Запись**	**Дата**	**Длина**	**Качество**	**Комментарии**	**Шум**
4	17390	https://chuvash.org/news/17390.html	95	https://ufile.io/jym6u	2017-11-10	0:01:05	✔		
5	17378	https://chuvash.org/news/17378.html	112	https://ufile.io/jp3m9	2017-11-14	0:01:29			
6	17407	https://chuvash.org/news/17407.html	78	https://ufile.io/ix3q7	2017-11-14	0:01:21			
7	17394	https://chuvash.org/news/17394.html	108	https://ufile.io/939h9	2017-11-14	0:01:34			
8	17362	https://chuvash.org/news/17362.html	92	https://ufile.io/xavds	2017-11-14	0:01:20			
9	14981	https://chuvash.org/news/14981.html	97	https://ufile.io/0gqxg	2017-11-14	0:01:17			high-amplitude white noise in the
10	14951	https://chuvash.org/news/14951.html	170	https://ufile.io/vlj9a	2017-11-15	0:01:55			high-amplitude white noise in the
11	17389	https://chuvash.org/news/17389.html	91	https://ufile.io/al7rd	2017-11-14	0:01:20			
12	14459	https://chuvash.org/news/14459.html	94	https://ufile.io/diwy5	2017-11-15	0:01:14			high-amplitude white noise in the
13	17393	https://chuvash.org/news/17393.html	97	https://ufile.io/qjp3g	2017-11-15	0:01:26			
14	16922	https://chuvash.org/news/16922.html	110	https://ufile.io/ismay	2017-11-18	0:01:28			
15	17289	https://chuvash.org/news/17289.html	89	https://ufile.io/da972	2017-11-15	0:01:20			
16	17290	https://chuvash.org/news/17290.html	100	https://ufile.io/urn36	2017-11-15	0:01:16			
17	17414	https://chuvash.org/news/17414.html	106	https://ufile.io/fo448	2017-11-15	0:01:16			
18	17409	https://chuvash.org/news/17409.html	114	https://ufile.io/nrw4h	2017-11-15	0:01:37			

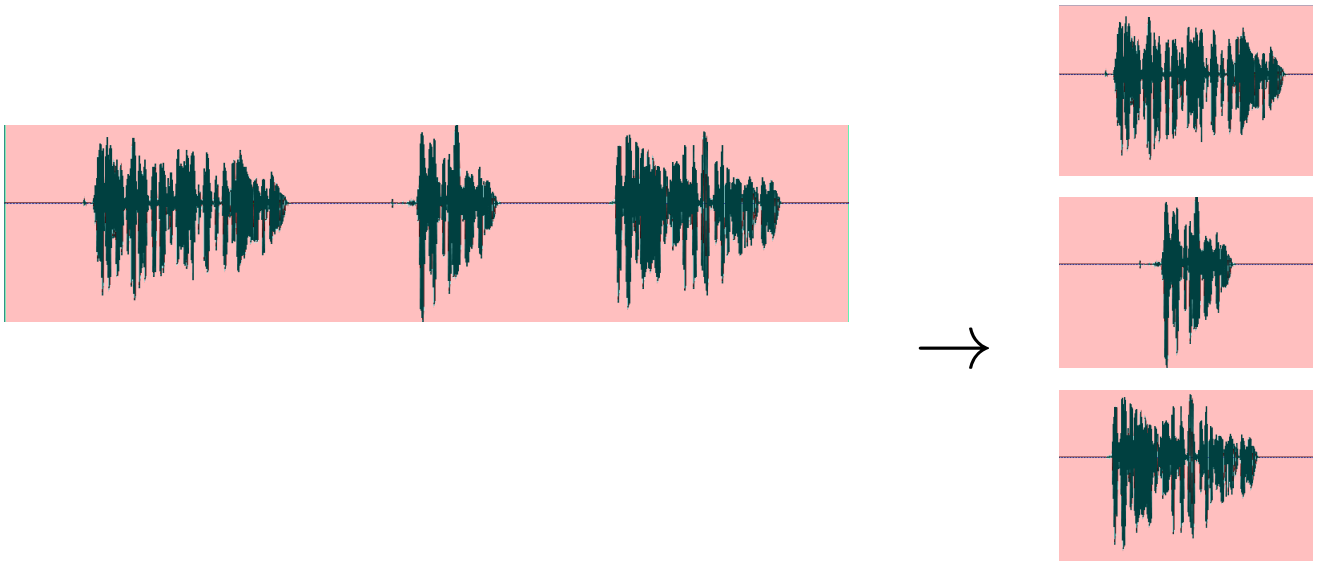
- Links files with text
- Some metadata
- Comments about quality, edits

Processing



- Use `sox` to make a silence profile
- Clean the audio using the profile
- Trim the first three seconds

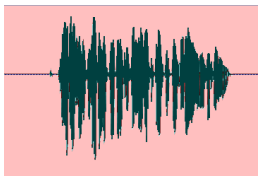
Processing



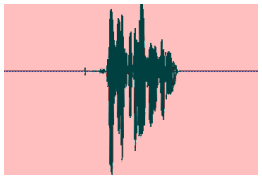
- Detect sentence spans using pyAudioAnalysis
- Split sentences using `sox`.

Processing

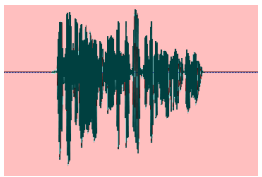
Result:



Çĕмĕрлесем хула тăрăх пĕр каяччă хăйне евĕр транспортпа çӳренине асăрханă.



Вăл гироскутера аса илтернĕ.



Иртен-çӳрен ăна видо ӳкерсе халăх те-телĕсене вырнаçтарнă.

- Text and sound sentence aligned corpus

Data

	Files	Sentences	Words	Minutes
Total:	299	2994	34648	384 (6h24m)
Postproc:	226	2090	21815	200 (3h20m)

Loss:

- 5 seconds per file (approx. 24m)
- 2 seconds per sentence (approx. 1h40m)

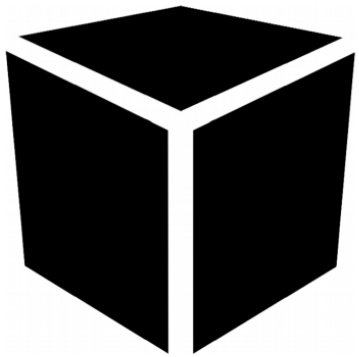
= 2h04m.

- Max: 4h20m
- minus approx. 25% of files
- $\rightarrow$ 3h20m

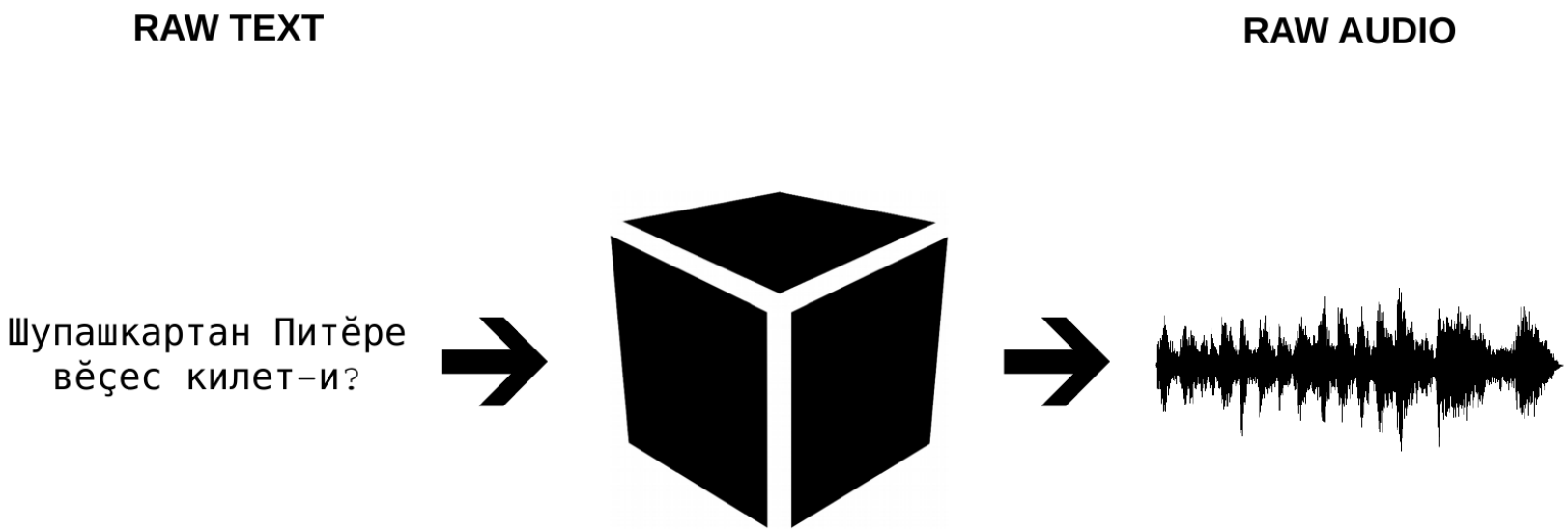
Modelling

Шупашкартан Питĕре вĕҫес килет-и?

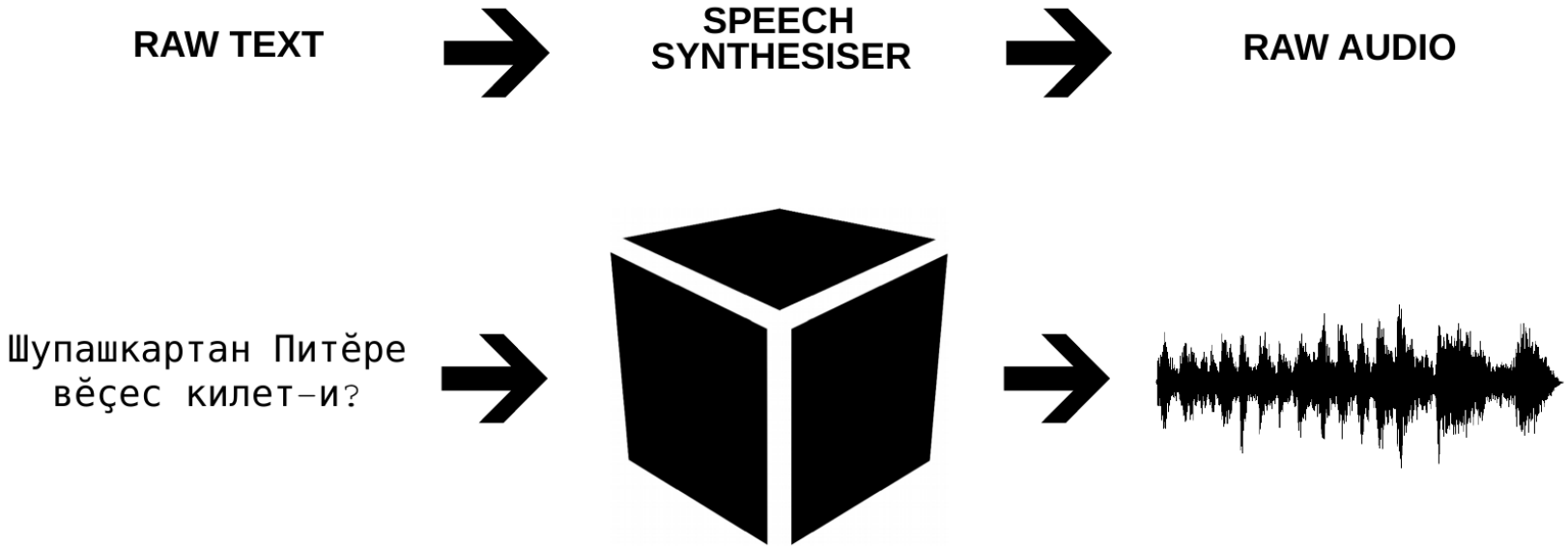
« Do you want to fly from Şupaşkar to St. Petersburg? »



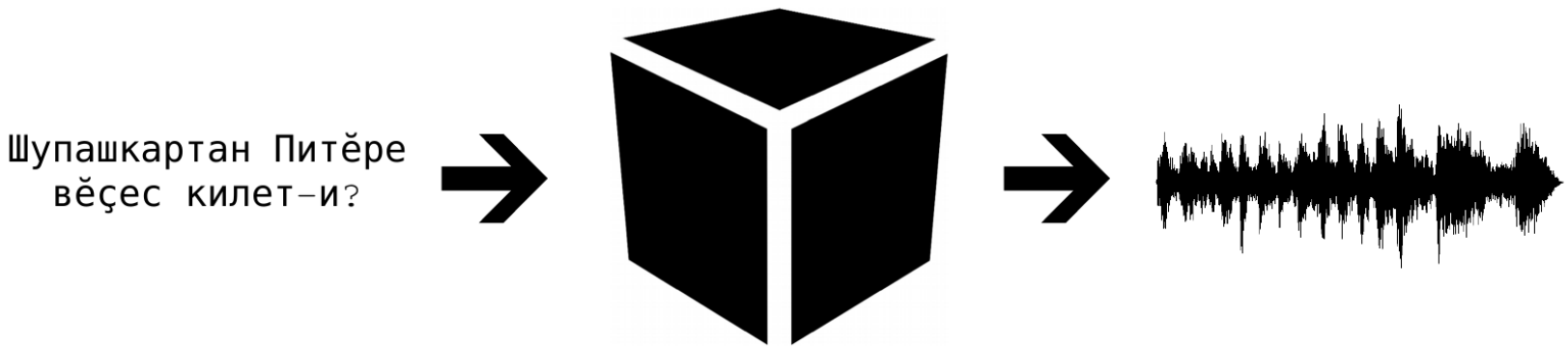
Modelling



Modelling

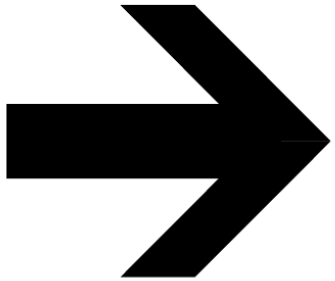

RAW TEXT
SPEECH SYNTHESISER
RAW AUDIO
Шупашкартан Питĕре вĕҫес килет-и?

Modelling



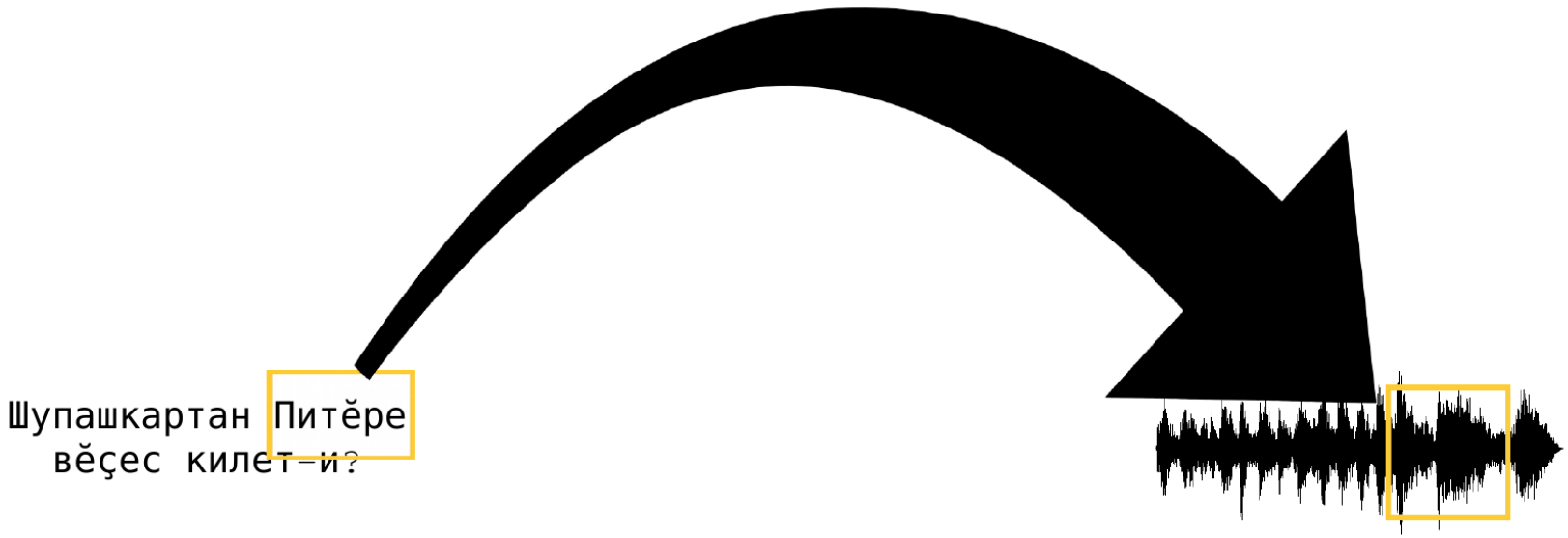
Modelling

Шупашкартан Питĕре вĕҫес килет-и?



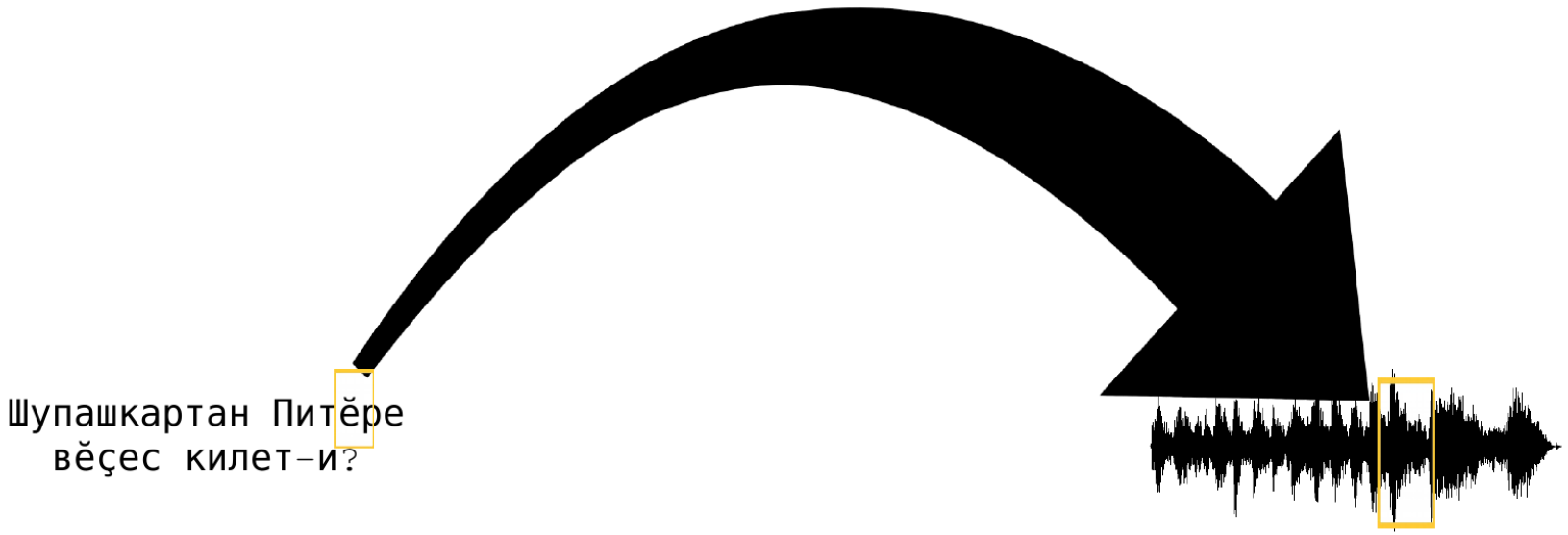
HARD

Modelling



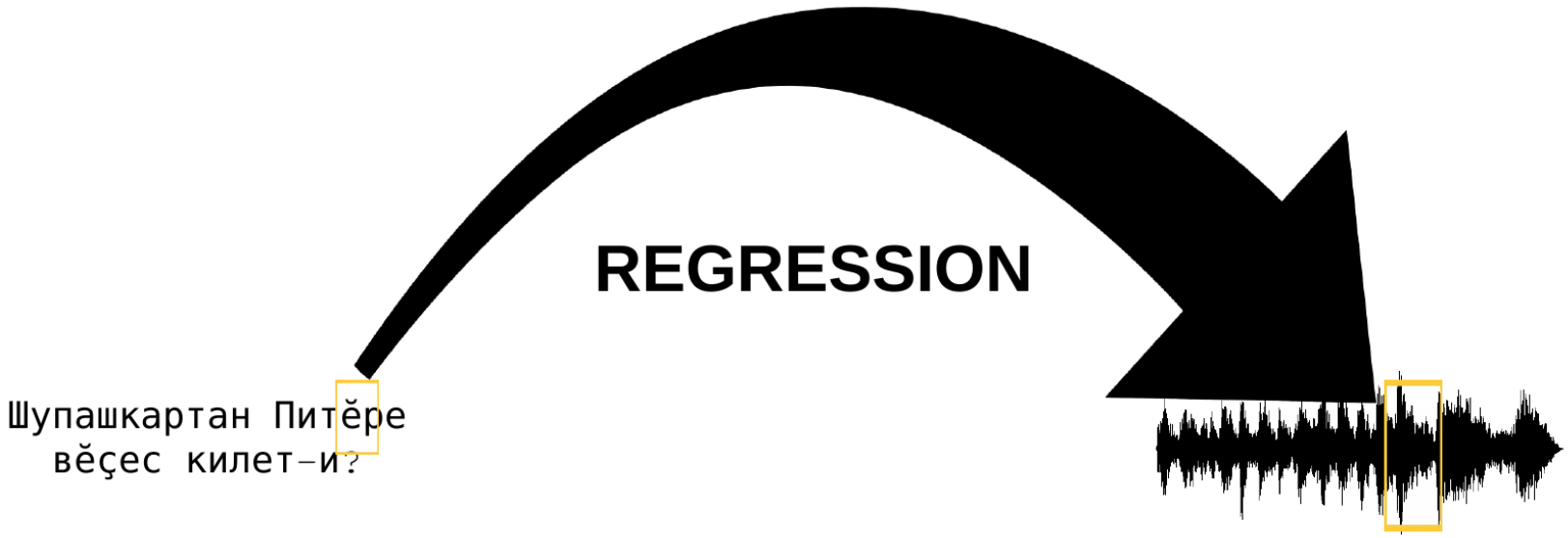
HARD

Modelling

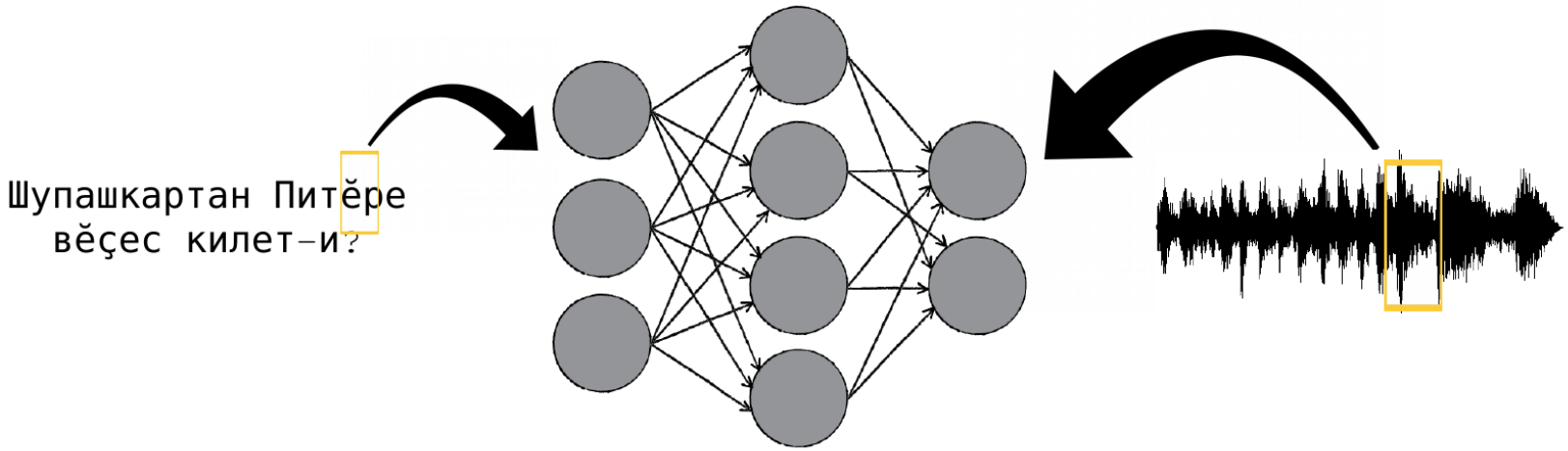


EASIER!

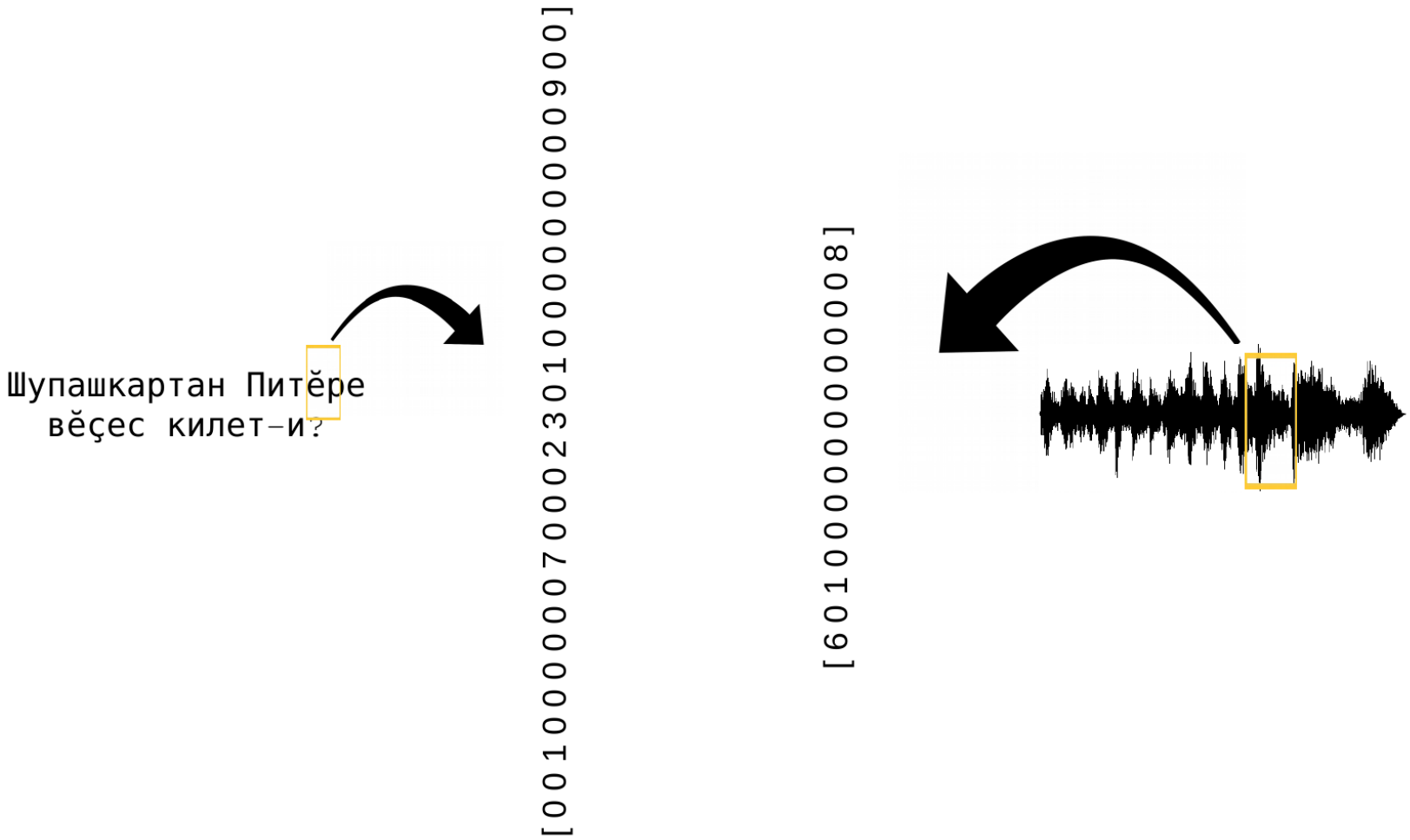
Modelling



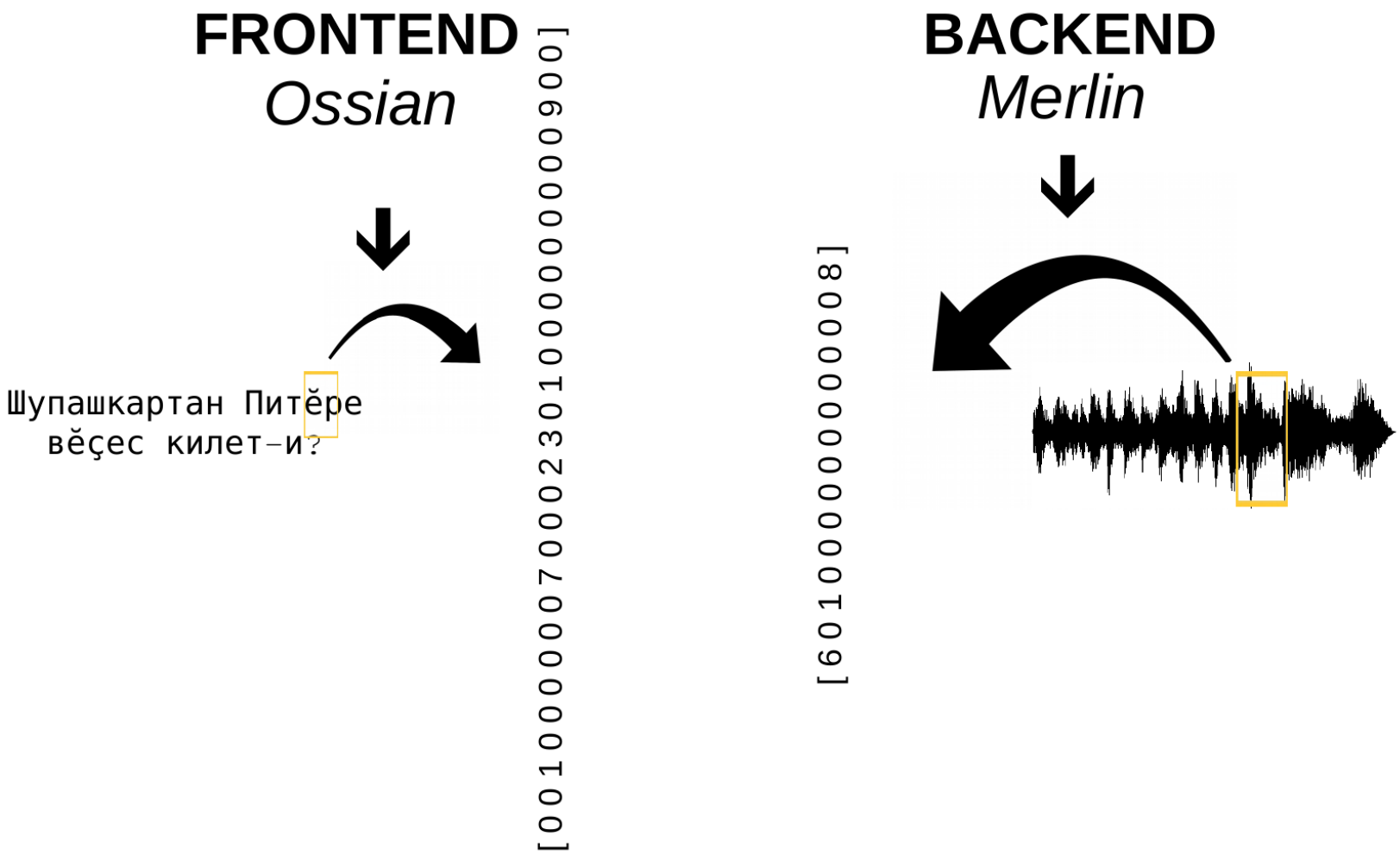
Modelling



Modelling


Шупашкартан Питĕре
вĕҫес килет-и?
[0010000007000230100000000009000]
[6010000000000008]

Modelling


FRONTEND
Ossian
Шупашкартан Питĕре
вĕҫес килет-и?
[0010000007000230100000000009001]
BACKEND
Merlin
[6010000000000008]

Frontend: Text features

Шупашкартан Питĕре вĕçес килет-и?

Letters

ш у п а ш к а р т а н п и т ĕ р е в ĕ ç е с к и л е т - и ?

Phonemes

ʂ u b a ʂ k a r d a n p i t ə r ɛ ʊ ə ʑ ɛ s k i l ɛ t _ i _

Words

[ʂ u b a ʂ k a r d a n] [p i t ə r ɛ] [ʊ ə ʑ ɛ s] [k i l ɛ t] [_ i _]

Parts of speech

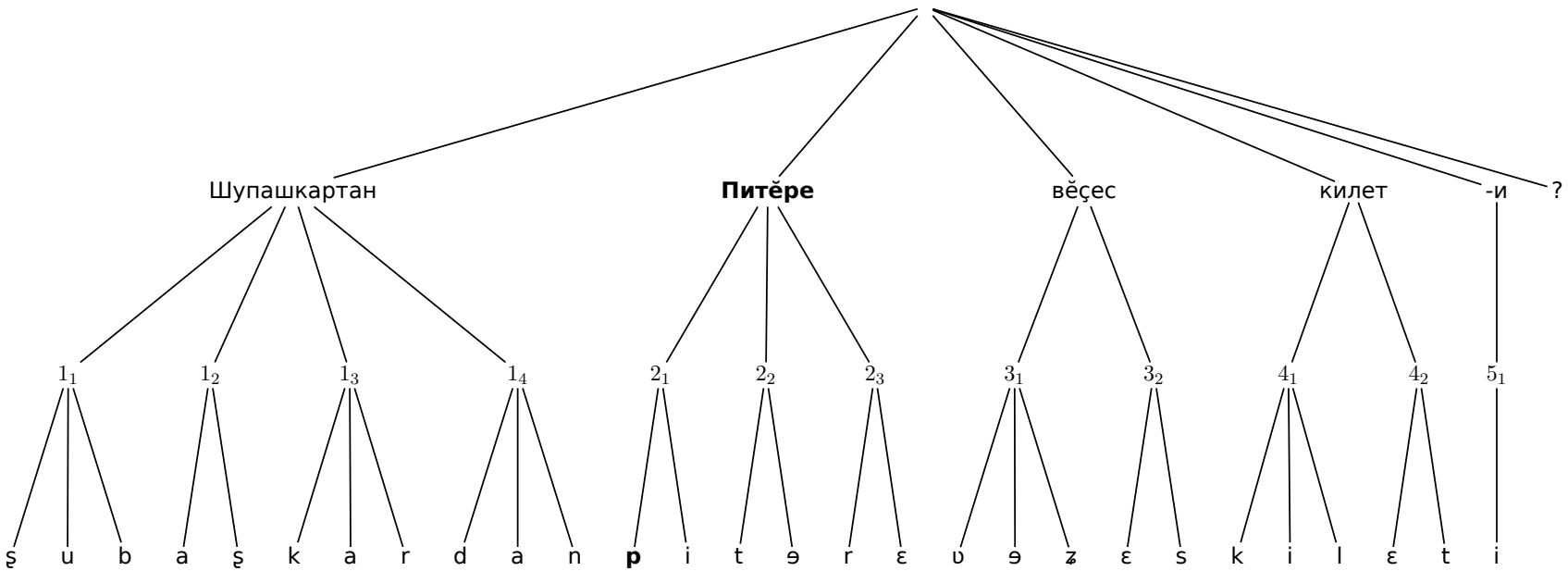
[PROPN ʂ u b a ʂ k a r d a n] [PROPN p i t ə r ɛ] [VERB ʊ ə ʑ ɛ s] [VERB k i l ɛ t] [AUX _ i _]

Syllables

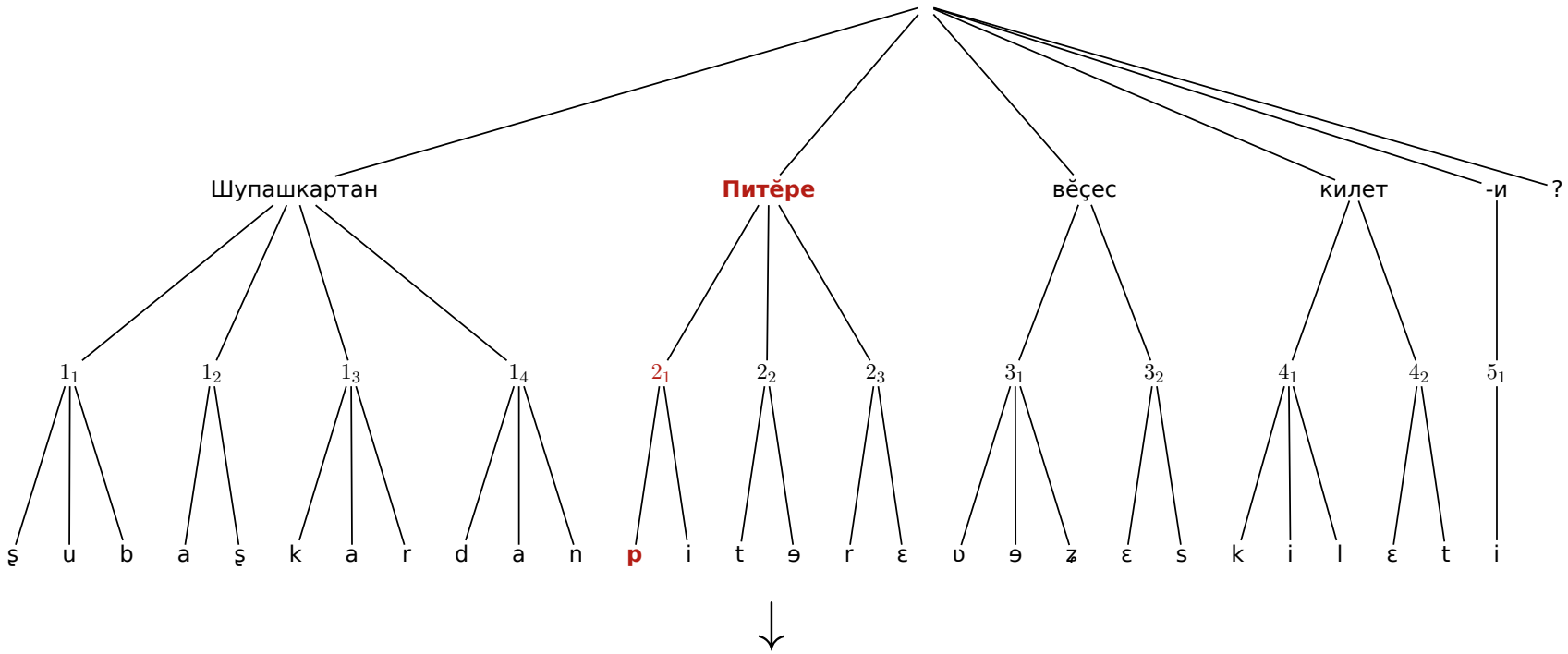
[PROPN [ʂ u] [b a ʂ] [k a r] [d a n]] [PROPN [p i] [t ə] [r ɛ]] [VERB [ʊ ə ʑ] [ɛ s]] [VERB [k i l] [ɛ t]] [AUX [_ i _]]

etc.

Frontend: Features to segments

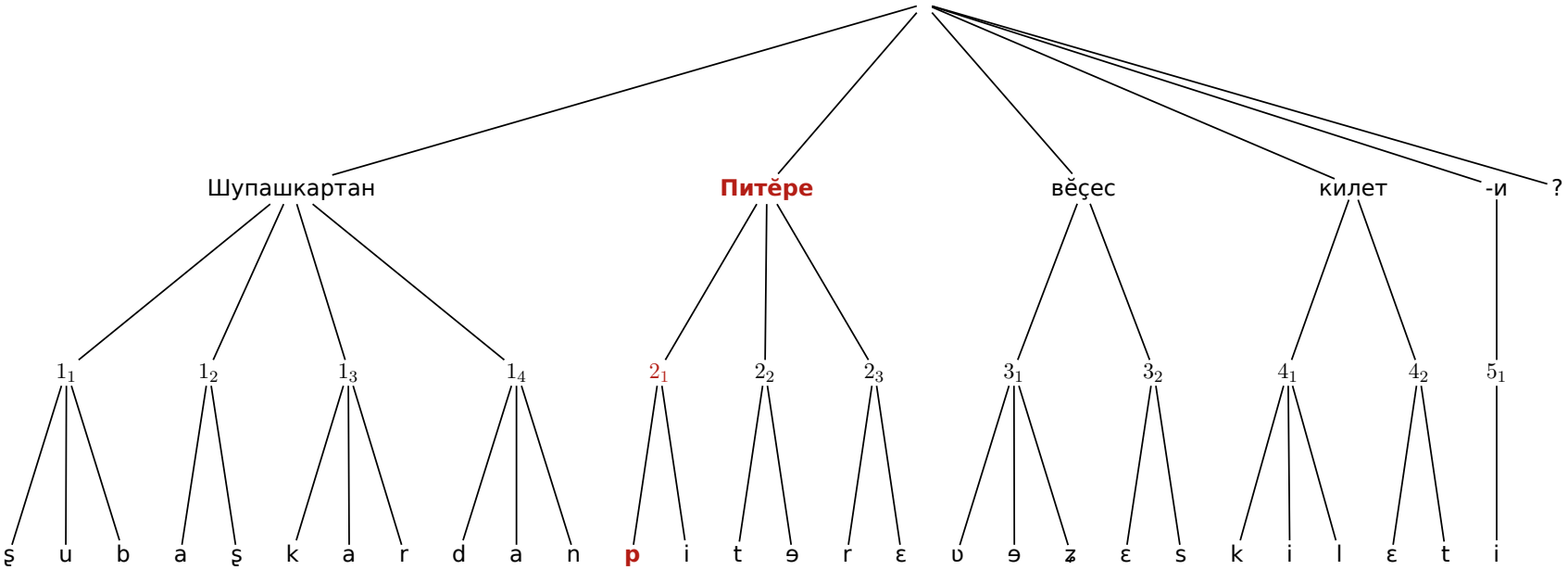


Frontend: Features to segments



c-1:n,c0:p,c+1:i,w0:Питĕре,w-1:Шупашкартан,w+1:вĕçес,s-1:dan,s+1:t□,…

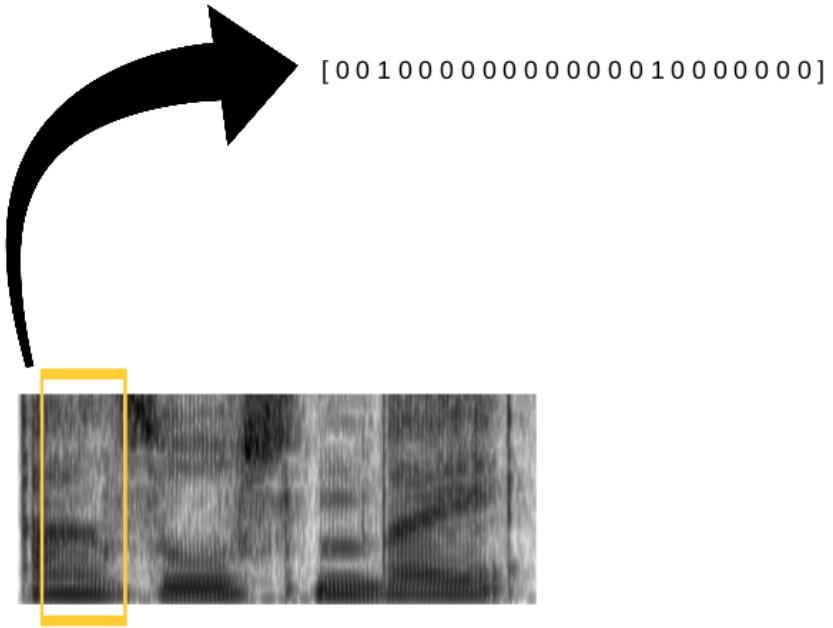
Frontend: Features to segments



c-1:n,c0:p,c+1:i,w0:Питĕре,w-1:Шупашкартан,w+1:вĕҫес,s-1:dan,s+1:t□,…

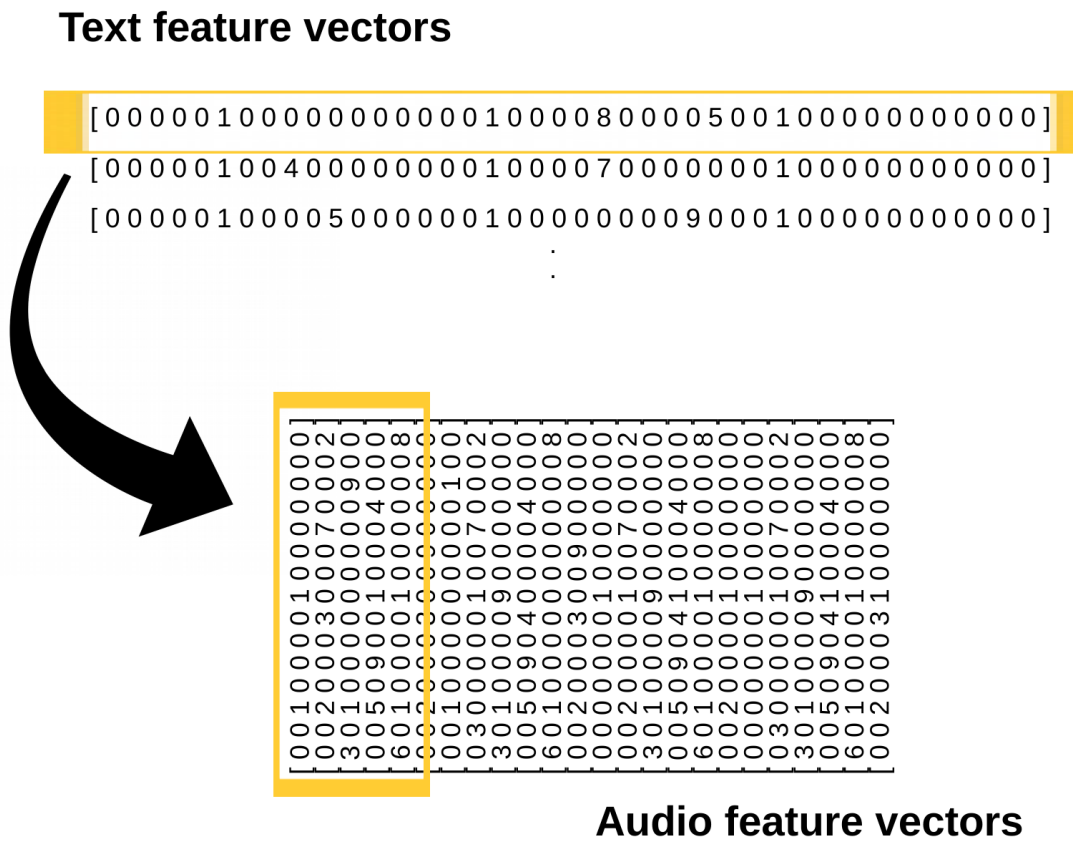
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0]

Backend: Audio features



- Each feature vector is a frame of audio (10ms)
- Features are amplitudes at frequency ranges

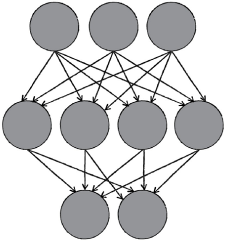
Alignment



- Unequal number of vectors for text and audio
- ...so just repeat them until they align

Training

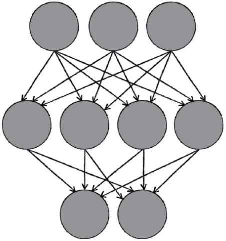
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]



[0 0 1 0 0 0 0 1 0 0 0 0 0 0 0]
[0 0 2 0 0 0 3 0 0 0 7 0 0 0 2]
[3 0 1 0 0 0 0 0 0 0 0 0 9 0 0]
[0 0 5 0 9 0 0 1 0 0 0 4 0 0 0]
[6 0 1 0 0 0 0 1 0 0 0 0 0 0 8]

Training

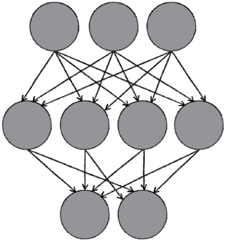
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]



[0 0 1 0 0 0 0 1 0 0 0 0 0 0 0]
[0 0 2 0 0 0 3 0 0 0 7 0 0 0 2]
[3 0 1 0 0 0 0 0 0 0 0 0 9 0 0]
[0 0 5 0 9 0 0 1 0 0 0 4 0 0 0]
[6 0 1 0 0 0 0 1 0 0 0 0 0 0 8]

Training

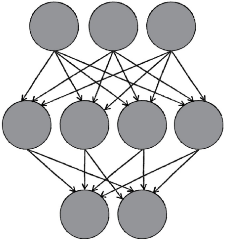
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]



[0 0 1 0 0 0 0 1 0 0 0 0 0 0 0]
[0 0 2 0 0 0 3 0 0 0 7 0 0 0 2]
[3 0 1 0 0 0 0 0 0 0 0 0 9 0 0]
[0 0 5 0 9 0 0 1 0 0 0 4 0 0 0]
[6 0 1 0 0 0 0 1 0 0 0 0 0 0 8]

Training

[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 8 0 0 0 0 5 0 0 1 0 0 0 0 0 0 0 0 0 0 0]



[0 0 1 0 0 0 0 1 0 0 0 0 0 0 0]
[0 0 2 0 0 0 3 0 0 0 7 0 0 0 2]
[3 0 1 0 0 0 0 0 0 0 0 0 9 0 0]
[0 0 5 0 9 0 0 1 0 0 0 4 0 0 0]
[6 0 1 0 0 0 0 1 0 0 0 0 0 0 8]

Evaluation

- 2 evaluators
- Scores 1 (very bad) — 5 (very good)

	Russian*	**Chuvash**
Intelligibility	5, 5	5, 3
Pronunciation	5, 5	5, 3
Stress	5, 5	4, 3
Intonation	2, 5	2, 3

- Intelligible
- Pronunciation and stress mostly ok
- Flat intonation

* Google TTS

Discussion

Future directions

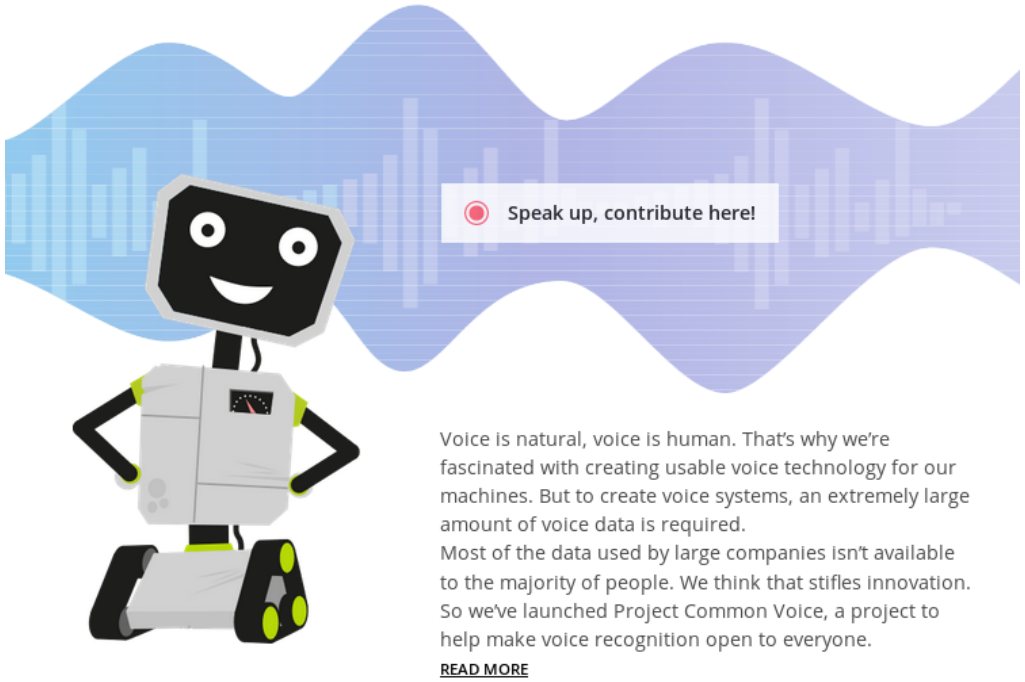
Speech synthesis

- Better linguistic features [POS, morphs, etc.]
- Fix segmentation of remaining 25%
- Numeral and abbreviation processing

Speech recognition

- Systems
  - Kaldi
  - Mozilla DeepSpeech
- Collecting more data
  - More volume (100s of hours)
    - How ?
  - More varied (different recording conditions)

Common Voice



- Web platform for collecting voice data
- Requires CC-0 — mostly impossible
- But…
  - Backend software is free/open-source

Ман çине вăхăт уйăрнишĕн тавах!

Merci beaucoup pour votre attention!

http://www.github.com/ftyers/Turkic_TTS