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
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»
<table>
<thead>
<tr>
<th>Synthesis (TTS)</th>
<th>Recognition (STT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Hours of audio</td>
<td>- 100s of hours of audio</td>
</tr>
<tr>
<td>- Studio quality</td>
<td>- Telephone quality</td>
</tr>
<tr>
<td>- Single speaker</td>
<td>- Many speakers</td>
</tr>
<tr>
<td>- Not processor intensive</td>
<td>- Processor intensive</td>
</tr>
</tbody>
</table>
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

STT

Intent parser

TTS

script

script

script

F. M. Tyers

Высшая школа экономики
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?

Hi, how can I help?

An Artificial Intelligence for Everyone
Project description
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 → 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
• 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
• 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.»

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

- 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
- **pyAudioAnalysis** – [https://github.com/tyiannak/pyAudioAnalysis](https://github.com/tyiannak/pyAudioAnalysis)

**Modelling:**
- **Ossian** – [https://github.com/CSTR-Edinburgh/Ossian](https://github.com/CSTR-Edinburgh/Ossian)
- **Merlin** – [https://github.com/CSTR-Edinburgh/merlin](https://github.com/CSTR-Edinburgh/merlin)
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
Data storage and organisation

- uploadfiles.io
- Files last 30 days
- Easy to use
Data storage and organisation

- Links files with text
- Some metadata
- Comments about quality, edits
• Use `sox` to make a silence profile
• Clean the audio using the profile
• Trim the first three seconds
• Detect sentence spans using pyAudioAnalysis
• Split sentences using sox.
Результат:

Ҫӗмӗрлесем хула тӑрӑх пӗр каяччӑ хӑйне евӗр транспортпа ҫӳренине асӑрханӑ.

Вӑл гироскутера аса илтернӗ.

Иртен-ҫӳрен ľана видо ўкерсе халӑх тетелӗсене вырнаҫтарнӑ.

• Text and sound sentence aligned corpus
### Data

<table>
<thead>
<tr>
<th></th>
<th>Files</th>
<th>Sentences</th>
<th>Words</th>
<th>Minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total:</strong></td>
<td>299</td>
<td>2994</td>
<td>34648</td>
<td>384 (6h24m)</td>
</tr>
<tr>
<td><strong>Postproc:</strong></td>
<td>226</td>
<td>2090</td>
<td>21815</td>
<td>200 (3h20m)</td>
</tr>
</tbody>
</table>

**Loss:**

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

= 2h04m.

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

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

Шупашкартан Питёре вёчес килет-и?
Modelling
Modelling
Modelling
Modelling

Шупашкarten Питёре вëсëс килет-и?

HARD
Modelling

шупашкертан Питере вёсес килет-и?

EASIER!
Modelling
[0 0 1 0 0 0 0 0 0 7 0 0 0 2 3 0 1 0 0 0 0 0 0 0 0 0 9 0 0]

[6 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8]
Modelling

FRONTEND
Ossian

Шупашкартан Питёре вёсес килет-и?

[001000007000230100000000900]

BACKEND
Merlin

[60100000000000109]

F. M. Tyers

Высшая школа экономики
Шупашкартан Питĕре вĕчес килет-и?

Letters
ш у п а ш к а р т а н п и т е р е в е ч е с к и л е т - и ?

Phonemes
ʂ u b a ş k a r d a n p i t ə r ɛ v ə z ɛ s k i l ɛ t _ i _

Words
[ʂ u b a ş k a r d a n] [p i t ə r ɛ] [v ə z ɛ s] [k i l ɛ t] [ _ i _]

Parts of speech
[PROPН ʂ u b a ş k a r d a n] [PROPН п i t ə r ɛ] [VERB в ə z ɛ s] [VERB к i l ɛ t] [AUX _ i _]

Syllables
[PROPН [ʂ u] [b a ş] [k a r] [d a n]] [PROPН [p i] [т ə] [r ɛ]] [VERB [v ə z] [ɛ s]] [VERB [k i l] [ɛ t]] [AUX [_ i _]]

etc.
Frontend: Features to segments

Шупашкетан

Питёре

вё̆́сес

килет

И?

F. M. Tyers

Высшая школа экономики
Frontend: Features to segments

c-1:n, c0:p, c+1:i, w0: Питёре, w-1: Шупашкертан, w+1: вĕçес, s-1: dan, s+1: ти, ...

F. M. Tyers
Frontend: Features to segments

Шупашкертан

Питĕре

вĕчес

килет

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]

F. M. Tyers

Высшая школа экономики
Each feature vector is a frame of audio (10ms)
Features are amplitudes at frequency ranges
Unequal number of vectors for text and audio
...so just repeat them until they align
Training
Training
Evaluation

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

<table>
<thead>
<tr>
<th></th>
<th>Russian*</th>
<th>Chuvash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligibility</td>
<td>5, 5</td>
<td>5, 3</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>5, 5</td>
<td>5, 3</td>
</tr>
<tr>
<td>Stress</td>
<td>5, 5</td>
<td>4, 3</td>
</tr>
<tr>
<td>Intonation</td>
<td>2, 5</td>
<td>2, 3</td>
</tr>
</tbody>
</table>

- 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)
- Web platform for collecting voice data
- Requires CC-0 — mostly impossible
- But...
  - Backend software is free/open-source
Ман çине вӑхӑт уйӑрнишӗн тавах!

*Merci beaucoup pour votre attention!*

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