ParlaSpeech-HR – a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus

Nikola Ljubešić, Danijel Koržinek,
Peter Rupnik, Ivo-Pavao Jazbec

ParlaCLARIN 3 workshop
Marseille, 20 June 2022

Overview

- Background data
- Dataset construction process
- Dataset description
- ASR experiments
- Future work

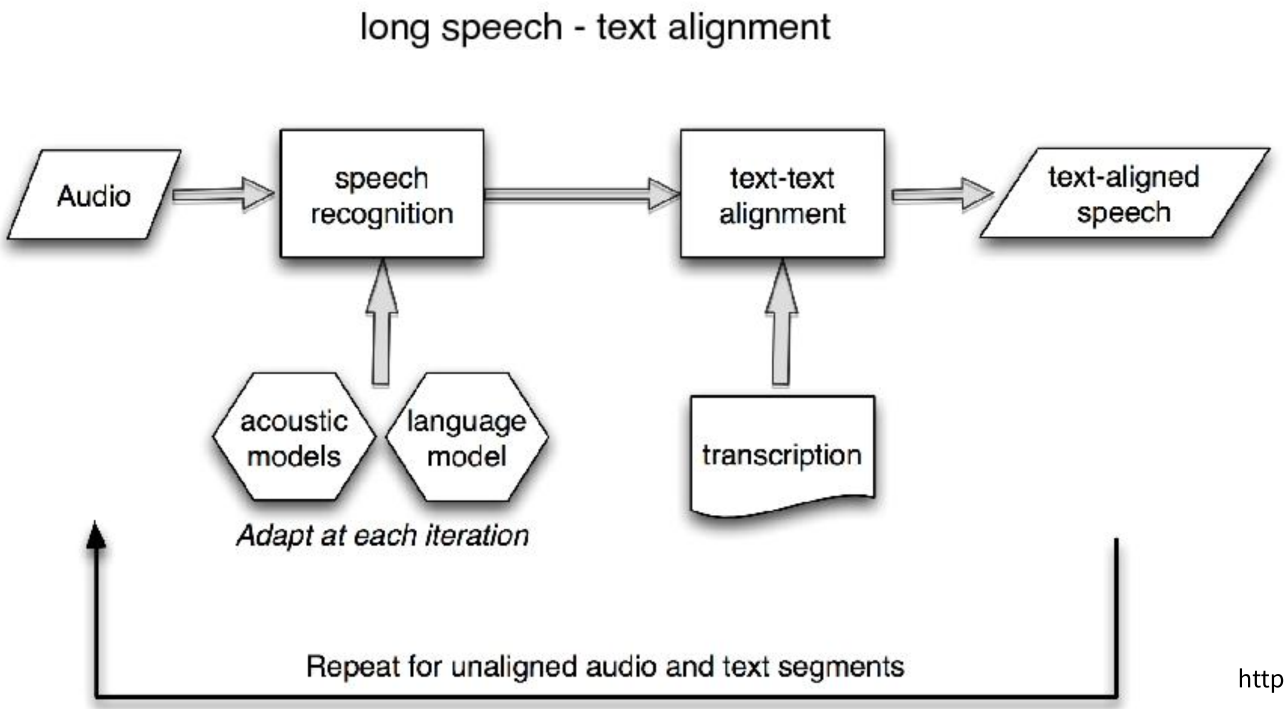
Background data

- Manual transcripts
  - ParlaMint 1 corpus (version 2.1)
  - Croatian is represented with data from one term (2016-2020), 20 million words
- Video / audio data
  - Harvested from the parliamentary YouTube channel
  - 2,419 hours of speech data

Dataset construction process (1)

- No open ASR model, very bad acoustic model
- Google speech-to-text has a reasonable ASR model for Croatian (word-error-rate 27.4%), and a $300 voucher
- Let us
  a. Transcribe some data with Google STT and align the automatic transcription with the manual transcription (Plüss et al. 2019)
  b. Learn an in-house model on the data from (a) and transcribe and align all available data

Dataset construction process (2)



long speech - text alignment

Audio

speech recognition

text-text alignment

text-aligned speech

acoustic models

language model

transcription

Adapt at each iteration

Repeat for unaligned audio and text segments

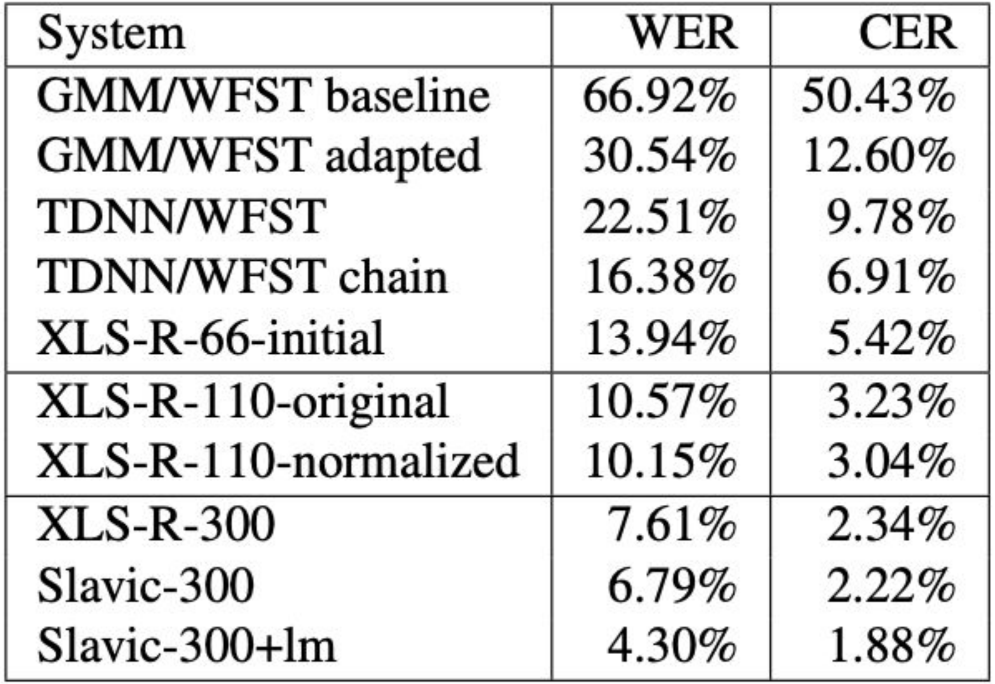
https://sail.usc.edu/old/software/SailAlign/

Dataset description

- Dataset obtained with Google STT – 66 hours
- Final dataset – 1,816 hours, 403,925 segments of 8-20 seconds in length (perfect for ASR training)
- Metadata on 309 speakers from ParlaMint – name, gender, age, party, party status
- train:dev:test split, in test three separate male and three female speakers
- http://hdl.handle.net/11356/1494 (CC-BY-SA 4.0)

ASR experiments

ParlaMint



System	WER	CER
GMM/WFST baseline	66.92%	50.43%
GMM/WFST adapted	30.54%	12.60%
TDNN/WFST	22.51%	9.78%
TDNN/WFST chain	16.38%	6.91%
XLS-R-66-initial	13.94%	5.42%
XLS-R-110-original	10.57%	3.23%
XLS-R-110-normalized	10.15%	3.04%
XLS-R-300	7.61%	2.34%
Slavic-300	6.79%	2.22%
Slavic-300+lm	4.30%	1.88%

Future work

- Speaker profiling benchmark - identity, gender, age, political orientation
- Datasets in other languages – Czech and Polish inside ParlaMint 2, but other, less resourced languages are very welcome to join the ParlaSpeech family!
- Multimodal corpora available through concordancers (video, gesture recognition?)

ParlaMint

<3 ?