Polish Read Speech Corpus for Speech Tools and Services

Danijel Koržinek, Krzysztof Marasek, Łukasz Brocki

Polish-Japanese Academy of Information Technology, Warsaw, Poland

October 28, 2016, Aix-en-Provence, France

About Clarin-PL

- B-type Clarin centre operating in Poland since 2013
- Run by teams from 6 Polish universities:
  - Wrocław University of Science and Technology
  - Institute of Computer Science PAS
  - PJAIT
  - Institute of Slavic Studies PAS
  - University of Łódź
  - Wrocław University
- It deals with various topics, including:
  - computer linguistics, social linguistics, language translation, language history and speech
- http://clarin-pl.eu

Speech resources at Clarin-PL

- Motivation:
  - lots of data used by HSS community exists in the form of audio
  - processing and analyzing this data is difficult and can be expensive
- This segment of the project is fully developed by PJAIT
- Consists of 2 main areas:
  - speech data
  - speech tools
- http://mowa.clarin-pl.eu

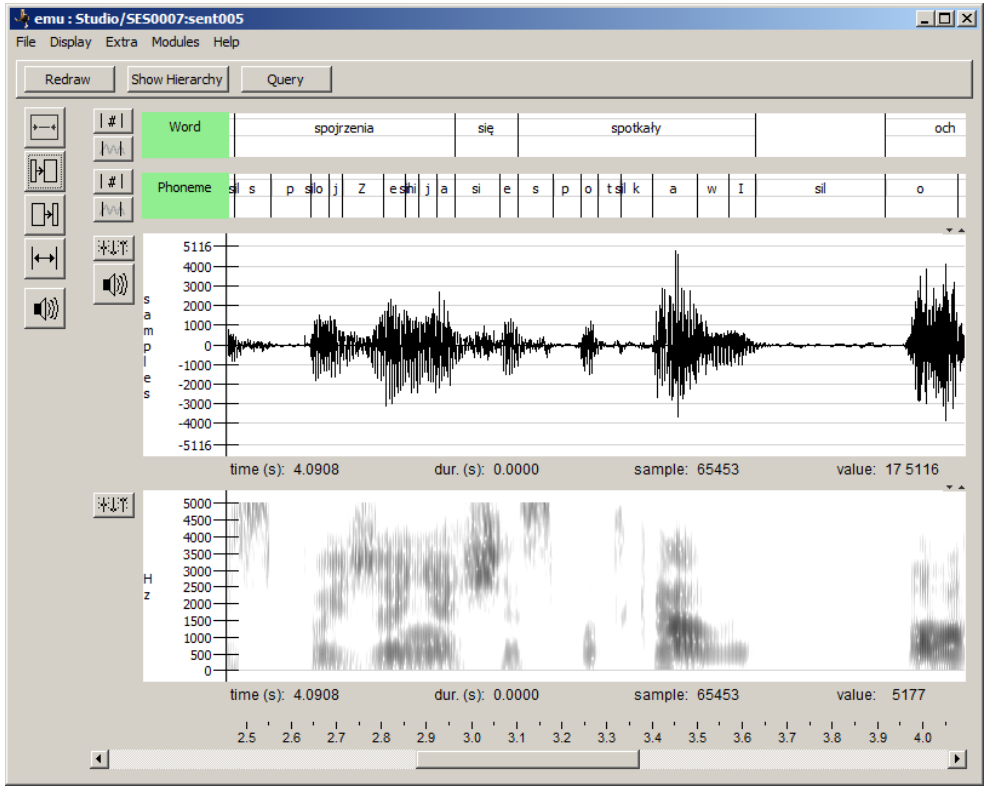
Speech data

- Speech corpora are expensive and hard to obtain, e.g.:
  - commercial: “CSLU”, “Speecon”, “GlobalPhone” and “Babel”
  - domain-limited: “Pelcra corpus of spontaneous speech” and “Spelling and NUmbers Voice database”
  - restricted due to copyright or other rules and limitation
- Our goal was to create a free general-purpose speech corpus

Polish general-purpose speech corpus

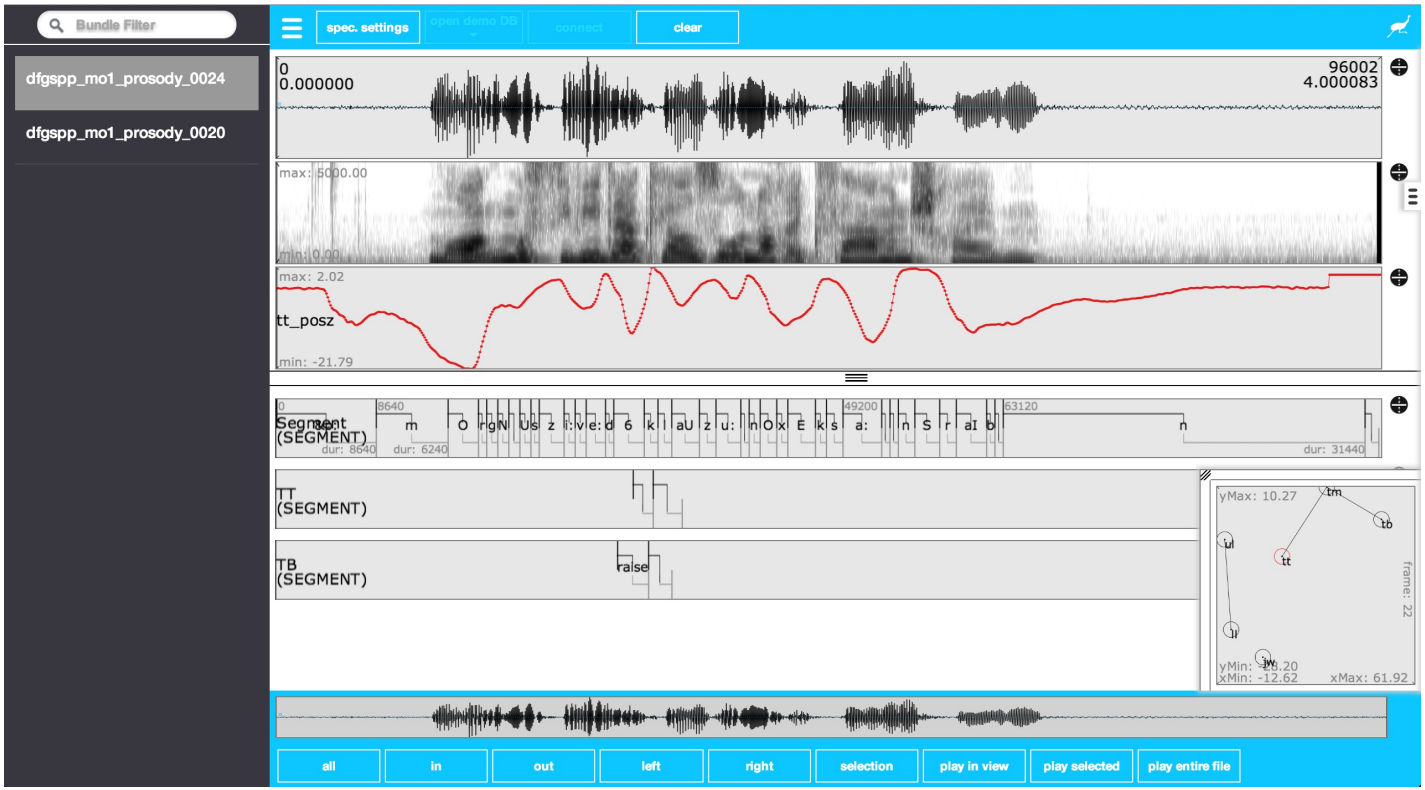
- We recorded and annotated ∼56h of studio quality read speech and ∼13h of telephone quality read speech
- The main purpose of these corpora is the development of speech processing tools
- It is available in two forms:
  - EMU database (currently old format)
  - Kaldi ASR baseline system
- It is available on a liberal license (CLARIN PUB+BY+INF+NORED)
- http://mowa.clarin-pl.eu/korpusy/

Emu database (old)



Emu database web service

- In development!



(source: http://ips-lmu.github.io/EMU.html)

Kaldi baseline results

WER %	experiment
30.06	mono
17.56	tri1
16.75	tri2a
15.75	tri2b
13.50	tri3b
13.10	tri3b-sp
12.88	tri3b-20k
12.41	tri3b-mmi
11.64	+wide beam
7.37	+large LM rescoring
3.23	oracle of wide beam
9.25	TDNN
5.91	+large LM rescoring
2.83	oracle
8.91	LSTM
5.78	+large LM rescoring
2.61	oracle

Kaldi baseline usage

- everything available on Github:

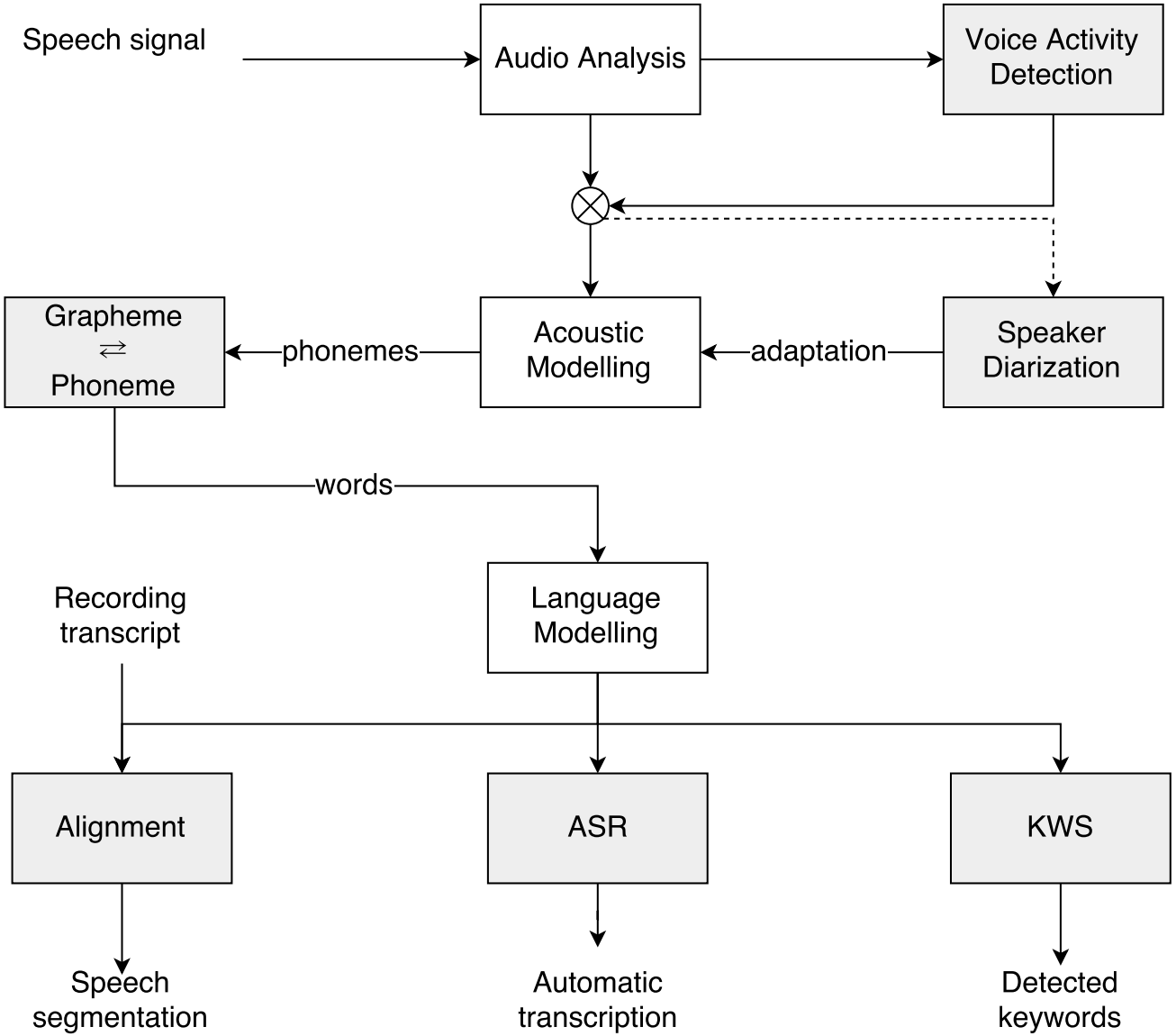
https://github.com/danijel3/ClarinStudioKaldi

- usage:
  1. download and install Kaldi
  2. `git clone ...` above repository
  3. modify path.sh and cmd.sh (if neccessary)
  4. `./run.sh`

Other speech data

- The recorded corpus, while useful for tool development, lacks certain features for the actual study of language
- We are currently working on other in-domain corpora:
  - PELCRA spontaneous speech corpus (with UŁ)
  - Polish Parliament (with UŁ and IPI PAN)
  - Kroniki - historical videos with news and current events (with University of Wrocław)

Speech tools diagram



Speech tools

- Other similar Clarin initiatives in other countries:
  - WebMAUS by LMU (speech segmentation)
  - AVATech by Max Planck Institute and Fraunhofer Institute (video/audio processing, speech segmentation, VAD and speaker diarization)
  - TTNWW (speech transcription services for Dutch)
- We developed speech tools available as web services:
  - Grapheme-to-phoneme conversion
  - Speech alignment
  - Speaker diarization
  - Voice activity detection
  - Keyword spotting
  - Speech transcription
- http://mowa.clarin-pl.eu/mowa

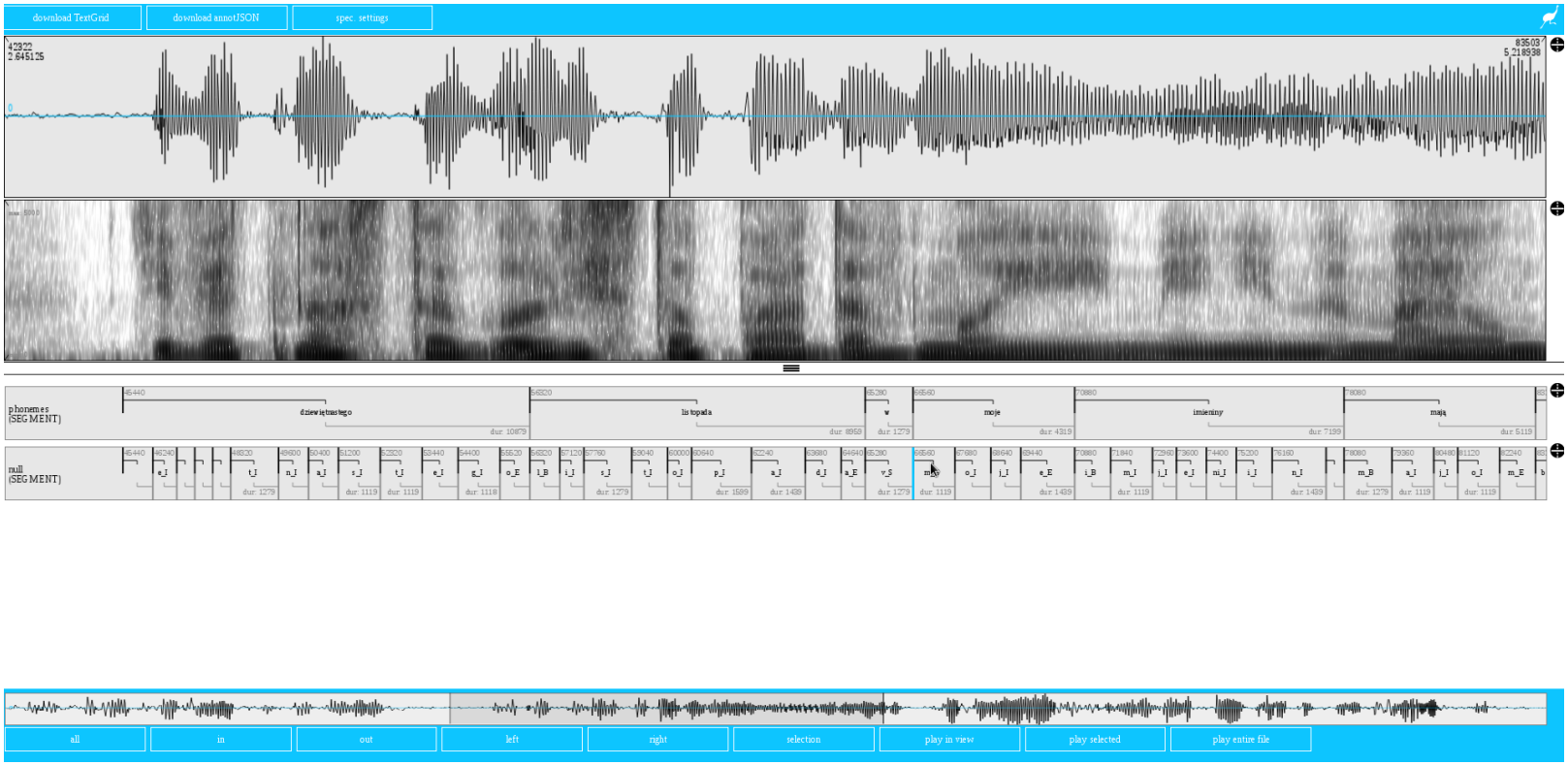
Grapheme-to-phoneme conversion

- Converting text from its orthographic into phonetic form
- Uses SAMPA phonetic alphabet
- Rule-based system
- Allows multiple word pronunciations
- http://mowa.clarin-pl.eu/transcriber

```
f S tS e b Z e S I ni e x S on S tS b Z m i f
t S tsi i ni e i S tS e b Z e S I n s t e g o
s w I ni e v u w g o p I t a p a ni e x S on S
tS u p o ts u S p a n t a g b Z en tS I v g on
S tS u
```

Text-to-speech alignment

- Given a transcription and an audio recording, we can calculate accurate alignment on word and phoneme level
- Also works on long audio (up to ∼30 minutes)



Voice activity detection

- "Naive" methods are easily deceived
  - thresholding, energy, 0-cross, running average,...
- Uses a trained acoustic model to reject non-speech events
  - knocks, noise, music, ...
- Difficulties with para-linguistic noise
- Uses a frame-based RNN model
- Has very high recall (>99%), but precision is still an issue (lots of noise can be misclassified as speech)
- Classification of non-speech was also attempted

Speaker diarization

- Multiple levels of speaker recognition:
  - speaker change detection
  - speaker diarization (← this was done)
  - speaker identificaiton
- Currently based on LIUM speaker diarization system
- Results are provided in the form of speech segmentation

Keyword detection

- Often we don't need a full transcript
- We can provide a list of keywords with the audio file and the system will generate a list of likely occurrences and their location
- It uses an ASR system with a general LM
- OOV words are modelled using syllables

Speech transcription

- Probably one of most sought after services
- Speech recognition works best when limited to a specific domain
- We provide a demonstration system for now, but would like to expand to specific domains useful in HSS research
- Based on the Kaldi toolkit for speech recognition

Selected applications

- Speech alignment/segmentation was used to annotate the Pelcra corpus of spontaneous speech
- Alignment was also used in the study titled “Respeaking - the process, competences and quality”
- Attempts were made to transcribe social science interviews

Future plans

- Additional corpora will be annotated and delivered on their respective platforms
- Usability improvements through integration with the EMU web platform
- Development of a transcription service aimed at HSS research
- Facilitation of cooperation with more partners in the HSS community

Contact

- Danijel Koržinek - danijel@pja.edu.pl
- Krzysztof Marasek - kmarasek@pja.edu.pl
- Łukasz Brocki - lucas@pja.edu.pl