CARINA – A CORPUS OF ALIGNED GERMAN READ SPEECH INCLUDING ANNOTATIONS

Hannes Kath[1,3] (hannes_berthold.kath@dfki.de), Simon Stone[1], Stefan Rapp[2], Peter Birkholz[1]
[1]Institute of Acoustics and Speech Communication, Technische Universität Dresden, [2]Fachbereich Informatik, University of Applied Sciences Darmstadt, [3]German Research Center for Artificial Intelligence (DFKI)

1. Motivation

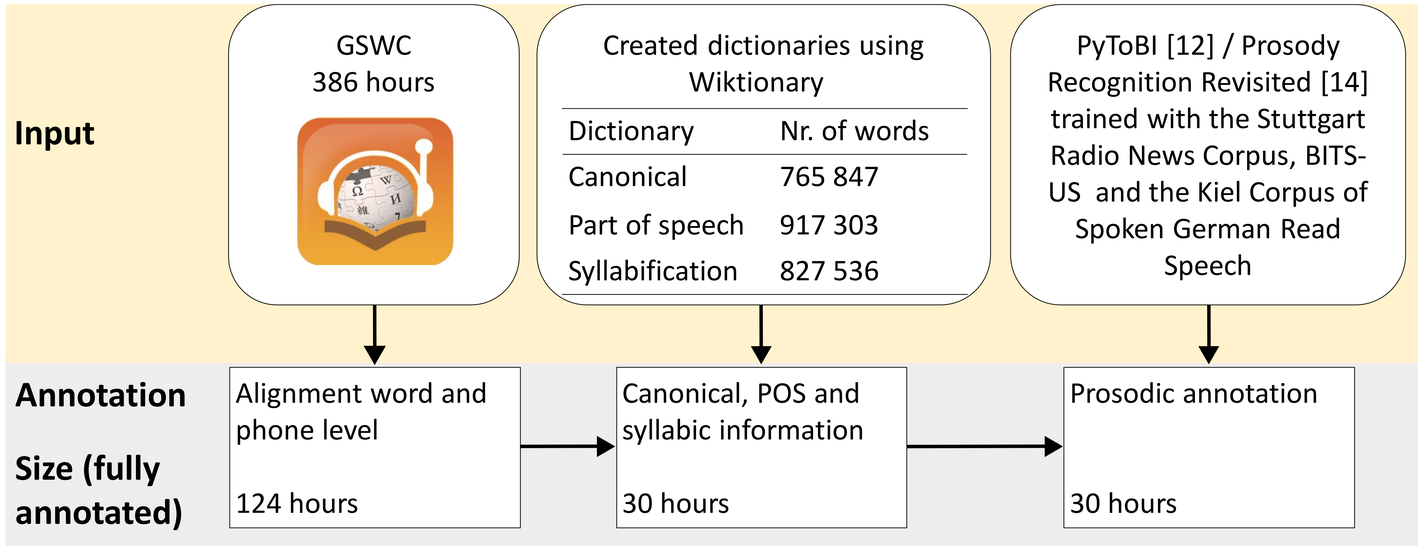
- Speech processing applications with neural networks require enormous amounts of data
- German speech corpora are not extensively annotated or of comparatively small size
- Manually annotated corpora are generally of higher quality than automatically created corpora, but usually expensive to create, therefore rarely under a free or permissive license
- CARInA uses a large amount of input data and strong selection criteria to form a carefully annotated, comprehensive German-language data set

2. Corpus Creation

German Spoken Wikipedia Corpus (GSWC) [6]

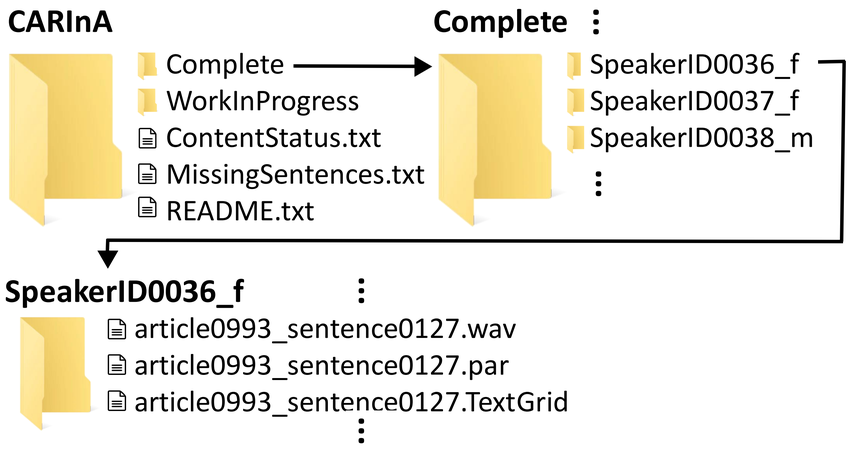
- CARInA[1] uses the speech material of the GSWC
- Read German Wikipedia articles on various topics
- Freely accessible
- 386 hours speech material
- 194 hours with complete sentences, of which the start and end samples are annotated (by MAUS and/or SailAlign)
- Grows over time (monitor corpus)
- 337 speakers (267 male, 36 female, 34 unspecified)

Automatic Annotation Pipeline



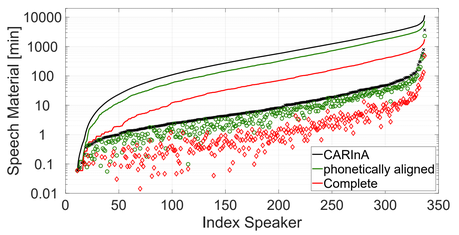
3. Data Set

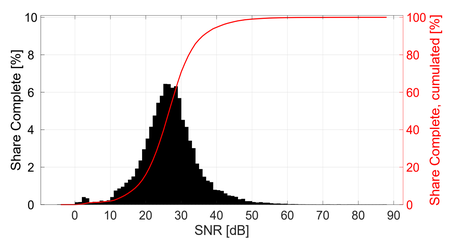
Corpus Structure



Properties

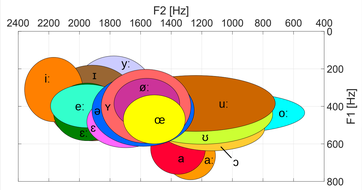
- Open source corpus
- 194 h speech material
- 124 h fully orthographically and phonetically aligned
- 30 h annotated on all speech levels
- 327 speakers (34 f, 259 m, 34 u)
- Average SNR: 26.8 dB





Validation

- Formant map for the subcorpus Complete
- Command word recognition system (CNN with 24 layers)



4. References

[1]http://dx.doi.org/10.25532/OPARA-144

[6] T. Baumann and A. Köhn, "The SpokenWikipedia Corpus collection: harvesting, alignment and an application to hyperlistening," Language resources and evaluation, vol. 53, pp. 303–329, 2016.

[12] M. Domínguez, P. Rohrer, and J. Soler-Company, "PyToBI: a toolkit for ToBI labeling under python," in Interspeech 2019. 2019, pp. 3675–3676, ISCA.

[14] S. Rapp, "Automatic labelling of German prosody," in International conference on spoken language processing 1998. 1998, pp. 1267–1270, ISCA.