CARINA – A CORPUS OF ALIGNED GERMAN READ SPEECH INCLUDING ANNOTATIONS

Hannes Kath¹,² (hannes_berthold.kath@dfki.de), Simon Stone¹, Stefan Rapp², Peter Birkholz¹

¹Institute of Acoustics and Speech Communication, Technische Universität Dresden, ²Fachbereich Informatik, University of Applied Sciences Darmstadt, ³German 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 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**

- Input: GSWC 386 hours
- Annotation: Alignment word and phone level 124 hours
- Created dictionaries using Wiktionary
  - Dictionary Nr. of words
    - Canonical 765 847
    - Part of speech 917 303
    - Syllabification 827 536
- PyToBI [12] / Prosody Recognition Revisited [14] trained with the Stuttgart Radio News Corpus, BIT-SUS and the Kiel Corpus of Spoken German Read Speech
- Grows over time (monitor corpus)
- 337 speakers (267 male, 36 female, 34 unspecified)

3. Data Set

**Corpus Structure**
- CARInA Complete WorkInProgress ContentStatus.txt MissingSentences.txt README.txt
- SpeakerID0036_f article0993_sentence0127.wav article0993_sentence0127.par article0993_sentence0127.TextGrid
- SpeakerID0037_f
- SpeakerID0038_m

**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

