TECHNISCHE UNIVERSITÄT DARMSTADT

Open acoustic models and speech data for German speech recognition

In the course of the BMBF project Dialog+, the LT and the Teleccoperation group have developed acoustic models for German distant speech recognition. These have been built with the open source software toolkits Sphinx and Kaldi. Unfortunately, German data resources needed to train such acoustic are rarely open source and easily accessible. We thus decided to record our own German speech data corpus, which we have now released under an open source license (CC-BY). Pretrained models and scripts to generate those are also available (see download links below) and are released under the same permissive CC-BY license.

The generation of this speech data corpus is supported by the BMBF project dialog+:

- Project description page
- Project homepage: dialogplus.eu

Summary of collected data (March 2015)	
Overall duration per microphone:	about 36 hours (31 hrs train / 2.5 hrs dev / 2.5 hrs test)
Count of microphones:	3 (Microsoft Kinect, Yamaha, Samson)
Count of wave-files per microphone:	about 14500
Overall count of participations:	180 (130 male / 50 female)

What is the difference to the freely available German Voxforge corpus?

- We have recorded all our speech data under controlled conditions: same room, same microphone distances, ...
- We recorded with three microphones in parallel. An additional signal was recorded with enabled beamforming and noise reduction (Microsoft Kinect).
- The data is curated, to reduce speaking errors and artefacts.

Downloads

- Complete speechdata-corpus 2014/2015 (wave&xml-files, tar.gz, ~17 GB, updated at 07/13/2015)
- Additional text resources for a German language model (gzip-txt, 8 million sentences, ~400 MB)
    - Automatically cleaned: German_sentences_8mil_filtered_maryfied.txt.gz
    - Raw: German_sentences_8mil.tar.bz2
- Sphinx AM: Feb/Mrz 2014 - v0.1 (approx. 40 participants)
- Sphinx AM: Feb-August 2014 - v0.2 (approx. 140 participants)
- Sphinx AM: Mai-2015 - v0.3 (approx. 180 participants)
- Sphinx AM: Oct-2015 - v0.4 (approx. 180 participants)
- Kaldi AM: Feb-March 2015 - v0.1 (approx. 180 participants - Download and instructions at github)

People

- Chris Biemann
- Dirk Schnelle-Walka (now at S1nn GmbH & Co KG)
- Stephan Radeck-Arneth (now at MaibornWolff GmbH)
- Benjamin Milde