Reorganized speech databases for Swedish ASR from Nordisk språkteknologi

The Norwegian Language Bank, July 2020

Introduction

In 2011, the Language Bank at the National Library of Norway received several speech resources for the Scandinavian languages from the bankruptcy estate of the company Nordisk språkteknologi (NST), among them several speech databases for dictation, ASR and speech synthesis. The speech databases for Swedish speech recognition were reorganized in 2013, and this document describes the reorganized version of these databases. General information about the speech databases can be found in the original documentation.

In addition to offering this database as a downloadable resource, you can also search and play the audio files on this page.

The original version of the database is available here.

The content of the databases

The speech databases consist of manuscript-read speech with associated annotations and metadata. The was originally divided into a training part, called *swe0467*, and a test part, called *swe0468*. Only *swe0467* is part of the reorganized database, but *swe0468* can be found in the original version. The metadata come in the form of JSON files, one file per recording session, while the audio is divided into one file per manuscript line, which often corresponds to a sentence.

In the original version of the material as it was handed over from NST, the files were organized in a specific folder structure where the folder names were meaningful (cf. the documentation). However, the file names were not meaningful, and there were also cases of files with similar names in different folders. This proved to be impractical, since users had to keep the original folder structure in order to use the data. The Language Bank has therefore renamed the files, so that the file names are unique and meaningful regardless of the folder structure. The original metadata files were in spl format. These have been converted to JSON format. The converted metadata files are also anonymized and the text encoding is

converted from ANSI to UTF-8. See below for more information on the metadata files and audio files.

The naming convention for metadata files is as follows:
language code + station number + x + informant code-recording date-time_original filename.json. As an example, the following information can be extracted from the file name *se10x016-08071999-1334_r4670016.json*:

- the language is Swedish
- the station number is 10[1]
- the informant number is 016
- the recording date is July 8 1999
- the recording time is 1.34 PM
- the original filename was *r4670016.spl*

The naming convention for the audio files is as follows:
language code + station number + x + informant code-recording date-time_original filename_channel number.json. The file name *se10x016-08071999-1334_u0016001-1.wav,* which is the audio file for one of the manuscript lines in the metadata file mentioned above, indicates that the original fine name was *u0016001.wav* and that the file contains the sound from channel 1. Two-channel audio files do not specify a channel in the file name.

The folder structure of the reorganized speech databases is as follows:

- *ADB_SWE_0467/* contains the metadata files for the training set
- *lydfiler_16_1/se* contains subfolders with names such as *se10x016-08071999-1334*, i.e. *language code + station number + x + informant code-recording date-time*. These contain all the channel 1 audio files with the specifications of the folder name
- *lydfiler_16_2/se* is organized as *lydfiler_16_1/se*, but contains the audio from channel 2
- *lydfiler_16_begge/se* is organized as the two preceding folders, but contains two-channel audio files

Metadata files

The metadata files are in JSON format. Each metadata file contains information about one recording session with one informant. Note that one recording session corresponds to several audio files, since the audio files are divided into one audio file per manuscript line. Fields in the metadata file indicate file names, folder paths, and character encoding provide information about the state of affairs in the original material from NST. As explained above, the character encoding has changed, and the folder structure is no longer needed to identify the correct file.

These are the fields of the metadata file:

- *info* contains metadata about the informant.

[1] Note that the station number in the file names does not always match the value of *station* in the metadata files. At the time of writing, it is unclear what this discrepancy is due to.

- *pid* contains the identifier for the recording session which is used in the updated filenames, i.e. *language code + station number + x + informant code-recording date-time.*
- *session* and *system* contain metadata about the recording session, original folder path and various technical information
- *val_recordings* contains the individual manuscript lines and metadata pertaining to these.
    - *text* contains the text of the manuscript line.
    - *file* contains the original file name of the audio file that renders the manuscript line. (The current file name is made up of the id in the *pid* field and the original file name, as explained above.)
    - *DST, NOI, QUA, SND, SPC, UTT* are fields pertaining to the validation of the recording. These fields are unused in most cases, containing only default values.
    - *t0, t1* og *t2* contain numbers attached to the manuscript lines in the original metadata files. It is unclear what purpose these numbers served.
    - *type* classifies the manuscript lines in different categories. The values often consist of a category identifier + a number, e.g. *CD20*, where *CD* indicates that the manuscript line is a sequence of numbers. Some type codes are more complex, such as ISp3, where *IS* is a category identifier for sentences. While the significance of the category identifiers is known to some extent, it is not clear what the other parts of the type codes signify. The following category identifiers were identified at the time of the reorganization of the databases in 2013:
        - *FF:* the text is a multi-word number
        - *IS:* the text is a sentence
        - *CD:* the text is a sequence of numbers
        - *pIW:* the text is a spelled word
        - *prIW:* the text is a name of a person (first name and last name)
        - *CIW:* the text is a proper name (a place or person)
        - *phIW* og *IW:* the text consists of one word

Audio files

The audio files in the original version of the speech database were in wav format with 16 kHz sampling frequency, 16 bit resolution and two channels (close and distant).[2] These files can be found in the folder *lydfiler_16_begge/se* in the new version of the database. When reorganizing the database in 2013, the two channels were also split using the software Sox. *lydfiler_16_1/se* contains wav files with the close channel, while *lydfiler_16_2/se* contains the audio from the distant channel.

[2] In the documentation attached to the original version of the database, it is claimed that the file format is headerless raw. However, the files in the database are in fact wav files.