

Acoustic database for Swedish

1. BASIC INFORMATION

1.1 Corpus composition

SpeechDat database (mobile phone and landline data for Norwegian, Swedish and Danish)
Telia (recordings for speech recognition made in office environments in the Stockholm area)
Recordings for speech recognition: ADB_OD_Swe.SWE database (recordings in office environment based on phonetically balanced manuscripts produced from sentences from the NST Swedish corpus)
Recordings for dictation: ADB_D_IBM-S database produced from recordings in office environment based on phonetically balanced manuscripts produced from news texts in the Swedish NST corpus Database of non verbal sounds marking hesitation

1.2 Representation of the corpora (flat files, database, markup)

Database

1.3 Character encoding

Not relevant

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, e-mail)

Arne Martinus Lindstad, ArneMartinus.Lindstad@nb.no

2.2 Copyright statement and information on IPR

Open Source

3 TECHNICAL INFORMATION

3.1 Data structure of an entry

Soundfiles in PCM/wav-format

Spl-loggfiler in txt

3.2 Corpora size (num. of tokens)

411 920 recordings

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

Monolingual, speech

4.2 The natural language(s) of the corpus

Swedish

4.3 Domain(s)/register(s) of the corpus

General speech

Phone conversations

4.4 Annotations in the corpus (if an annotated corpus)

Not relevant

4.5 Intended application of the corpus

Speech recognition, dictation, speech synthesis

4.6 Reliability of the annotations (automatically/manually assigned) – if any

No annotation, but the database has been validated by language assistants under the supervision of group leaders.

5 RELEVANT REFERENCES AND OTHER INFORMATION

<http://www.nb.no/spraakbanken/tilgjengelege-ressursar/taledatabasar>