In this document, $dirMain=/net/aistaff/aki.

## Building an acoustic model

In [Schuppler, Ernestus, Scharenborg, Boves, 2011], Barbara used ECSD (Ernestus Corpus of Spontaneous Dutch) to train acoustic model. She mentioned that CGN doesn't have enough data to train acoustic models for specific sounds. In our study, CGN was used for the preliminary test. The performance is not very high, but forced alignment seems working most of the time.

The processing is written in $dirMain /src/acoustic\_model

The following data is used.

* Wav: /net/corpora/CGN\_2.0.3/data/audio/wav
* Transcription: /net/corpora/CGN\_2.0.3/data/annot/text/fon

First of all, .fon files are copied and unzipped.

Each wav file is splitted into each utterance with transcription (.txt) based on its fon file ($dirMain/src/acoustic\_model/split\_wav.py).

The transcription (.txt) of splitted wav files are converted into HTK label files ($dirMain/src/acoustic\_model/script2label.py). In this steps, sentences which include '#', '[' and ‘]’ and '-' are removed.

Fn001023\_162-809.lab is removed, because it has transcription which is not listed in the phone list.

$ find ./ -type f -print | xargs grep 'hoge'

Fa2fon.py

ʉ -> u (gekust)

When the word is not found in the dictionary, please use ipa2pronvar.sh to generate pronunciation variation.

## Evaluation data

Overview

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sentence ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Groningen and Drenthe | 174 | 160 | 117 | 55 | 173 | 76 | 53 | 55 | 76 | 77 |
| Limburg | 214 | 201 | 146 | 91 | 214 | 108 | 90 | 90 | 109 | 109 |
| Oost Overijssel-Gelderland | 163 | 157 | 108 | 55 | 164 | 70 | 54 | 56 | 72 | 72 |
| Total | 551 | 518 | 371 | 201 | 551 | 254 | 197 | 201 | 257 | 258 |
| Number of words | 19 | 9 | 10 | 11 | 11 | 19 | 21 | 5 | 7 | 10 |

## Initial analysis

Azure ML

## ToDo

* Acoustic model should be trained with ECSD
* HDMan using two sources

## Classification

*Make sure that you train and test each model on the basis of balanced data (set per sentence to the maximum possible). So for sentence 1: 3 x 163 sentences per region (divided into 130 x 3 sentences train and 33 x 3 test), and for sentence 10: 3 x 72 sentences (58 train and 14 test per region). Besides reporting the results on the basis of the test set, please also determine results use 10-fold cross validation on the training set. I.e. divide the training set into 10 sets of 90% + 10% data (e.g., for sentence 1: 10 sets of 117 x 3 training files, and 13 x 3 test files; all 10 test sets constitute the total 130 x 3 sentences in the full training set). Then you can train on the 117 x 3 files, and test on the 13 x 3 files. You do this for each of the 10 splits, and create a summed confusion matrix (in which all 130 x 3 documents will be assigned a cluster). This then also allows you to calculate the accurracy, etc. By having both results on the test set and 10-fold cross validation on the training set, we get a better idea of the actual performance. Now the current test on the basis of 10% of the data (also trained using non-balanced data) likely gives a wrong picture of the performance.*

Mail from Martijn on 2017/10/20

For sentence 1-10, 10 balanced datasets are made. In total,