An Evaluation of Unsupervised Acoustic Model Training for a Dysarthric Speech Interface

Oliver Walter[1], Jort F. Gemmeke[2], Vladimir Despotovic[1], Bart Ons[2], Reinhold Haeb-Umbach[1] and Hugo Van hamme[2]

[1]Department of Communications Engineering - University of Paderborn
[2]ESAT - PSI Speech Group - KU Leuven

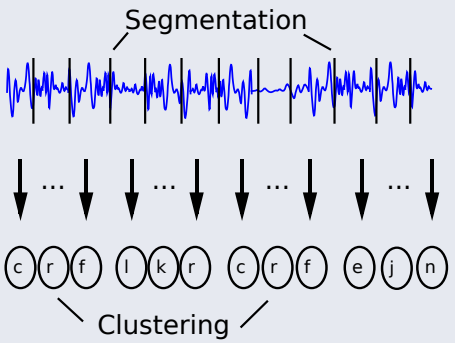
September 16, 2014

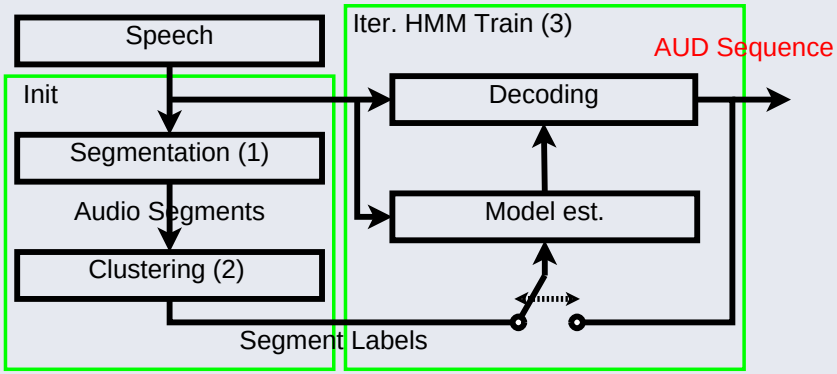
Acoustic model training

Acoustic model training for speech impaired persons

- Speaker independent training performs poorly for speech impaired persons
- Speaker dependent training of acoustic models for improved performance
- **Here:** Unsupervised learning → transcription and spoken words are unknown
- Framewise: Vector quantization, Gaussian mixture models, posteriorgrams

Acoustic units: basic building blocks of sequences of speech frames





- Three steps:
  1. Segmentation of the speech signal at changepoints
  2. Clustering of similar segments into acoustic units
  3. Iterative HMM training of the acoustic models for each acoustic unit

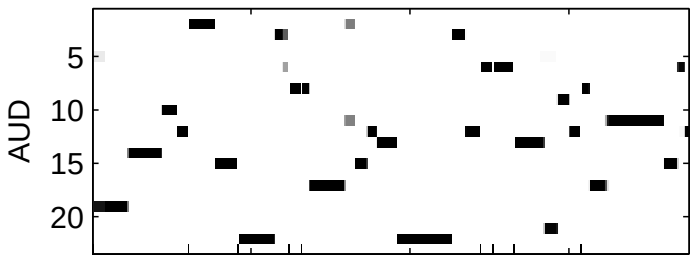
Examples for acoustic units

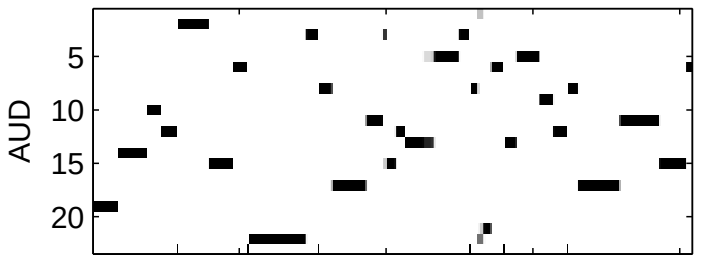
Acoustic unit sequences

- Two utterances of "ALADIN Hoofdeinde op stand 1" spoken by one speaker:
  AJ AE AA AC B AF F BJ C H H AH AB AF AC AD BJ C AC F F AD E I AC H AH AB AF F
  AJ AE AA AC B AF F BJ C H AH AB AF AC AD E C H BB F AD E I AC H AH AB AF F

Posteriorgram representation

- Using the acoustic models (HMMs) of the acoustic units:





⇒ Acoustic units deliver a consistent representation of similar utterances

Evaluation: Speech interface

Training of speech interfaces should be as simple as possible

- User speaks with his own words
- No restrictions on the words to be used for a certain command
- Only a semantic frame description per utterance is provided
  - Example: "Hoofdeinde op stand 1" $\Rightarrow$ Hoofdeinde 1
- Evaluation: Use AUD sequences and posteriorgrams as input to the word finding