Beyond symbolic level in transcriptions

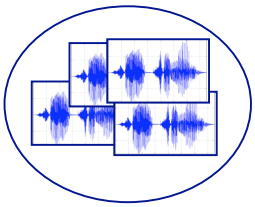
Michele Gubian
Radboud University Nijmegen
The Netherlands

SOUNDTOSENSE® Project 11

Transcribing corpora

- Phoneticians are often required to make yes/no decisions on presence/absence of segments
  - E.g. in Dutch, in word final /ən/, /n/ is often deleted (‘*lopen*’ -> /lopə/)
  - In careful speech /n/ is sometimes present (40% in read speech)
  - There are hard to decide cases
- Qualitative judgement
  - Listening
  - Spectrogram reading
- Quantitative judgement
  - Agreement among transcribers (voting)
- Can we shed some light in this problem with automatic quantitative tools?
  - Search for evidence in the signal for clear-cut dichotomy, like /ən/ vs. /ə/

Pattern analysis

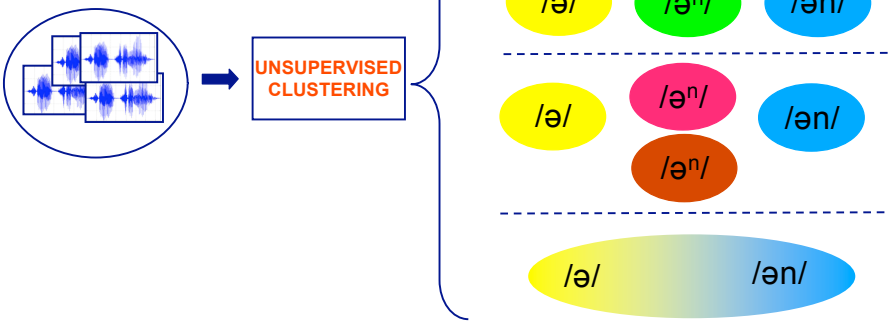


PATTERN ANALYSIS TOOL

Structure description

- The input set contains audio tracks of word final /ən/ in Dutch
- No phonetic labeling is included
  - "let the data speak"
  - Unsupervised clustering
  - The revealed structure may or may not provide evidence for a clear-cut dichotomy like /ən/ vs. /ə/

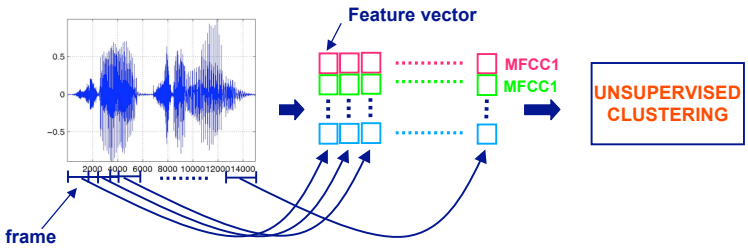
Possible outcomes



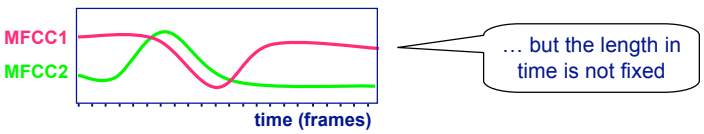
How to do that?

- Unsupervised clustering tools require an input of fixed dimension
  - but speech segments have variable duration
  - the same problem in ASR is addressed using HMMs
  - but HMMs are supervised models, and not for clustering
- Non-relevant information should be filtered out
  - partly solved by feature extraction
  - in ASR, Mel Filter Cepstral Coefficients (MFCC) are used
  - we can use them as well

Feature extraction



- To obtain …



Open problems

- We have to describe feature trajectories with a fixed size descriptor (vector, matrix…)
  - "Stretching and squeezing" trajectories might not be ok because information about their duration would be destroyed
  - Time series compression schemes might help
  - E.g. SAX (Symbolic Aggregate approXimation)
- We want to preserve the order of feature events
  - E.g. a peak in MFCC1 before a valley in MFCC2
  - Feature sets other than MFCC may relate feature events to articulation events
  - A set of articulatory features extracted from the speech signal using neural networks is under study

Clustering

- Unsupervised clustering is "blind"

- Clusters have to be cross-checked with known factors
  - Gender, age, …
- After that, cross-checking with manual transcriptions will show how information automatically extracted from the signal correlates with binary decisions taken by the phonetician
  - /ən/ vs. /ə/

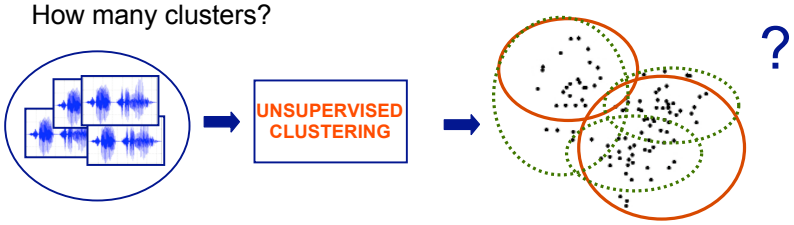
And yes, results will come soon…

Thank you!

Appendix

Clustering (2)

- How many clusters?



- "Siimple" clustering algorithms require to choose manually the number of clusters
  - E.g. k-means (k is the number of clusters)
- Several techniques exist to get to a "reasonable" cluster number
  - Empirical ones
  - Based on regularization

How to do that? (2)

- Audio data are segmented using an ASR
  - forced alignment using HTK and manually verified transcriptions
- Left and right context of /ən/ are chosen to be 'safe'
  - /{t,d} ə n # {p,k,t}/
  - This will provide reliable start and end points for each sample (landmarks)