Advances in
Regional Accent Clustering in Swedish

Giampiero Salvi
giampi@kth.se

Outlook

- **Introduction**
- **Method**
- **Data**
- **Results**

Introduction

- **aim: analysis of regional pronunciation variation on large data sets ($\sim$5000 speakers)**

Introduction

- aim: analysis of regional pronunciation variation on large data sets (∼5000 speakers)
- how? Automate part of the process with data mining techniques

Introduction

- aim: analysis of regional pronunciation variation on large data sets ($\sim$5000 speakers)
- how? Automate part of the process with data mining techniques
- inspiration: analysis of L2 speakers (Minematsu and Nakagawa, 2000)

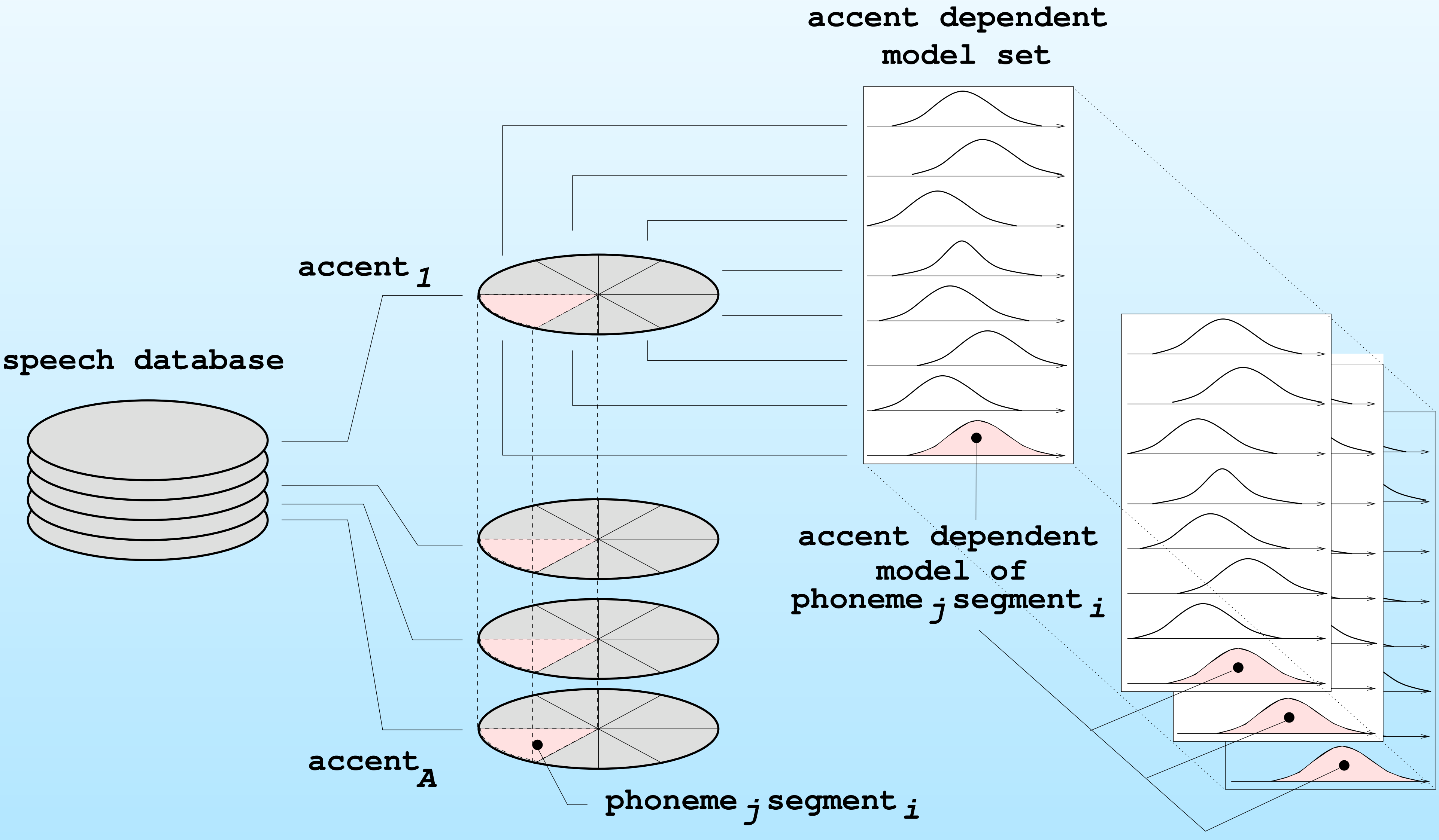
Introduction

- aim: analysis of regional pronunciation variation on large data sets (∼5000 speakers)
- how? Automate part of the process with data mining techniques
- inspiration: analysis of L2 speakers (Minematsu and Nakagawa, 2000)
- **previous work:**
  - **Analysis of accent variation of single phonemes (Salvi, 2003a)**
  - **Use of accent information in ASR (Salvi, 2003b)**

Method (1)

- **first use ASR (Automatic Speech Recognition) techniques to collect statistics for each phoneme**
  - **divide database in A subsets depending on accent region**
  - **extract acoustic features at fixed time intervals**
  - **build accent dependent monophone models with one distribution per state**
- **result is a pdf for each phoneme $ph_1, \cdots, ph_P$, subsegment $s_1, \cdots, s_S$ and accent region $r_1, \cdots, r_A$**

Statistics with ASR



Statistics with ASR (cont.)

- **advantages**
  - do not need phonetic transcriptions
  - procedure can be automated and reproduced identical elsewhere
  - easy to deal with large databases

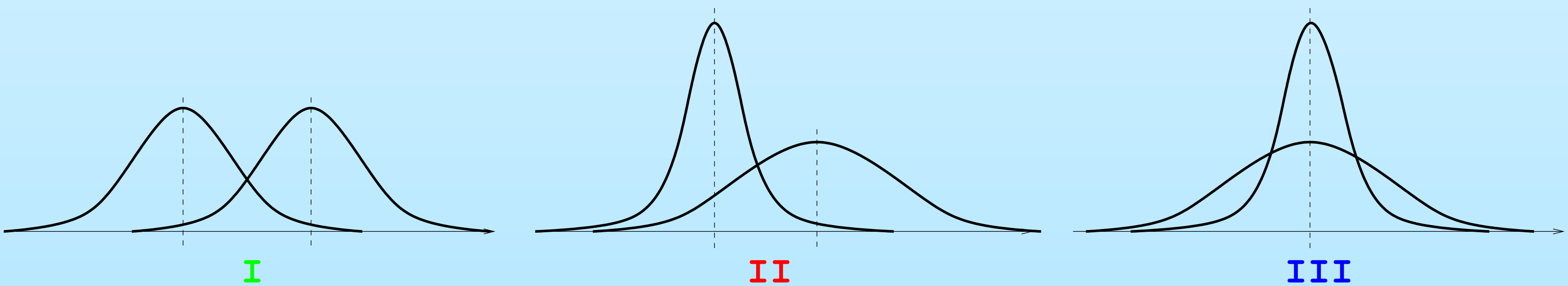
Statistics with ASR (cont.)

- advantages
  - do not need phonetic transcriptions
  - procedure can be automated and reproduced identical elsewhere
  - easy to deal with large databases
- disadvantages
  - pronunciation model based on dictionary (canonical)
  - harder to spot mistakes (if database is not clean)
  - suprasegmental (prosodic) features hard to include

Method (2)

- **Analyse differences between groups by comparing distributions**
  - **metric based on Bhattacharyya distance**

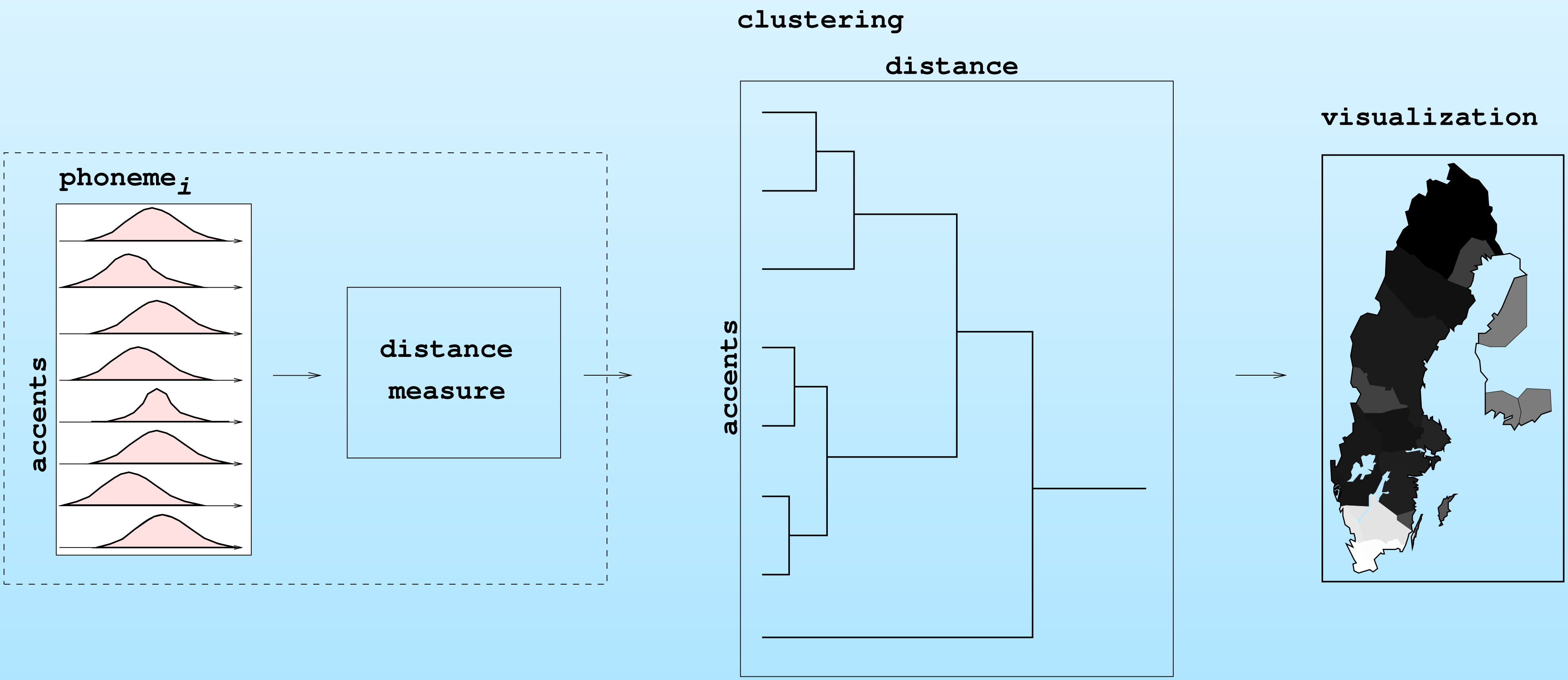
$$D_{\mathsf{bhatt}}(\Theta_1, \Theta_2) = \overbrace{\underbrace{\frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1)}_{\text{I}} + \underbrace{\frac{1}{2} \ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}}_{\text{III}}}^{\text{II}}$$



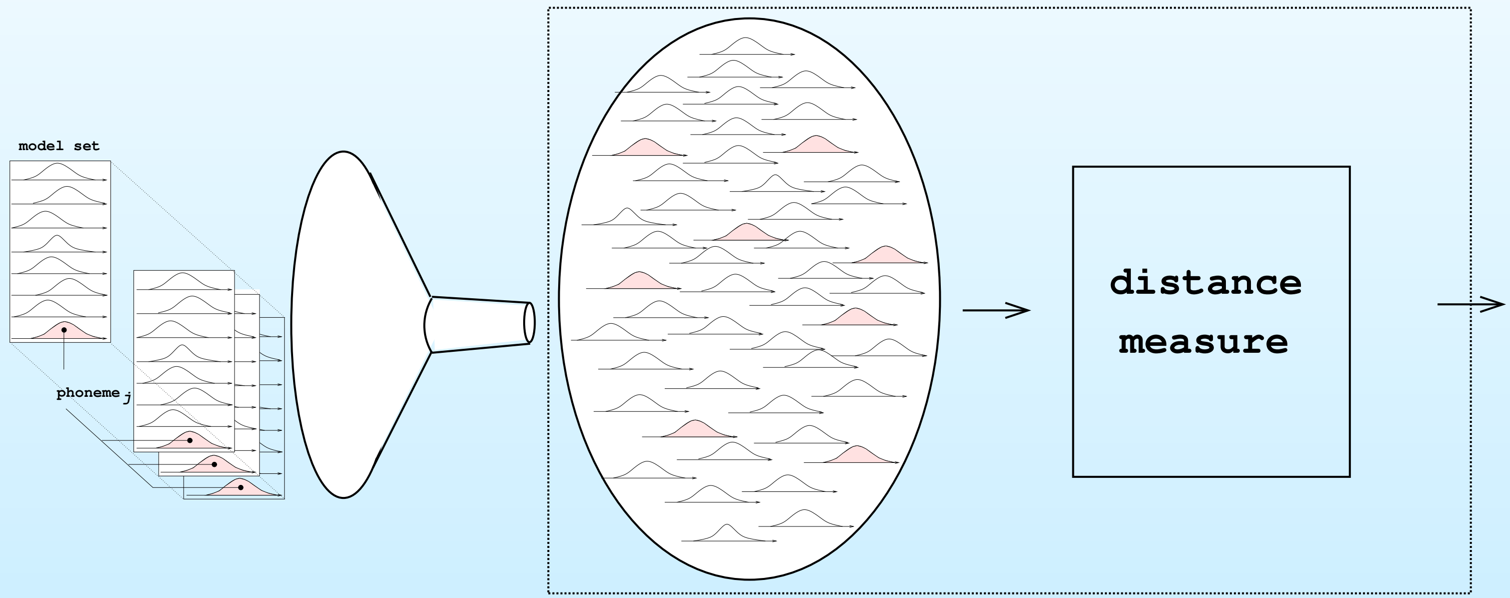
- **use agglomerative hierarchical clustering to interpret the data**

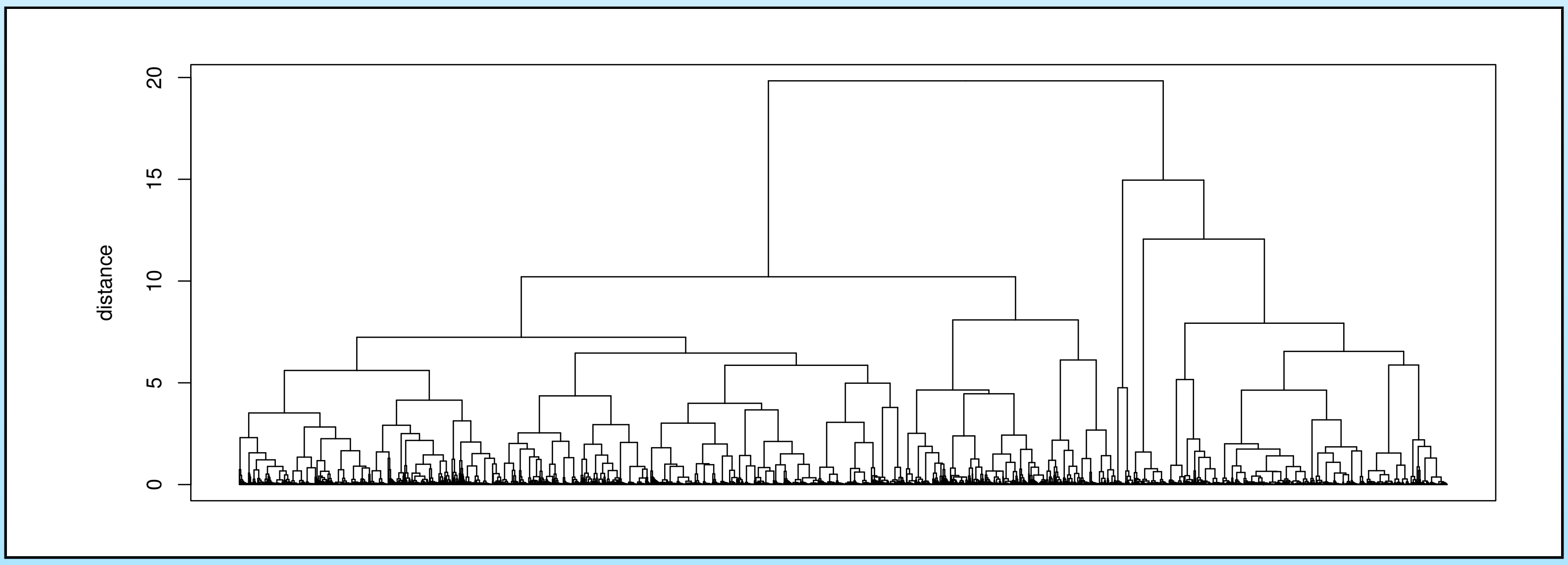
Analysis in previous studies

- In (Salvi, 2003a,b)
  - consider each phoneme independently
  - merge initial/middle/final subsegments



Analysis in this study





Analysis in this study (cont.)

- **advantages:**
  - **let allophones from different phonemes cluster together**
  - **enable observation of more general groups (consonants, vowels...)**
  - **study the initial, middle and final part of each phoneme separately**

Analysis in this study (cont.)

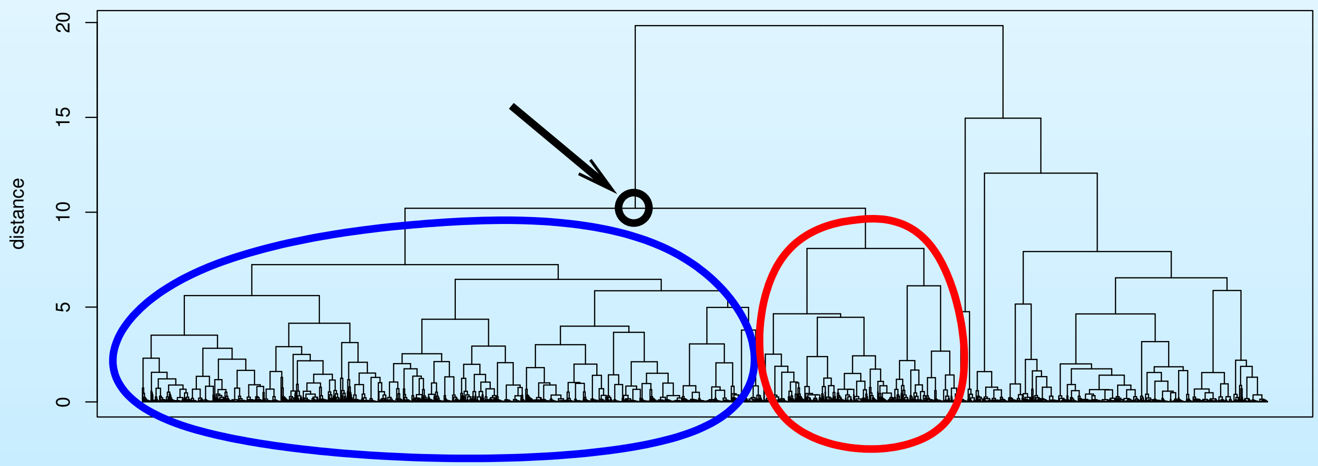
- advantages:
  - let allophones from different phonemes cluster together
  - enable observation of more general groups (consonants, vowels...)
  - study the initial, middle and final part of each phoneme separately
- disadvantages:
  - the clustering tree becomes huge
  - problems of visualisation

Data

- Swedish SpeechDat FDB5000
- 5000 speakers recorded over the telephone line
- 270 hours of recordings (including silence)
- 10msec spaced Mel frequency cepstrum coefficients $c_0, \ldots, c_{12}$
  $+$ 1st order differences $d_0, \ldots, d_{12}$
  $+$ 2nd order differences $a_0, \ldots, a_{12}$
- total of 96.803.850 data points (39 dim vectors)
- 20 accent regions $\times$ 46 phonemes $\times$ 3 subsegments $=$ 2760 distributions

Method (3)

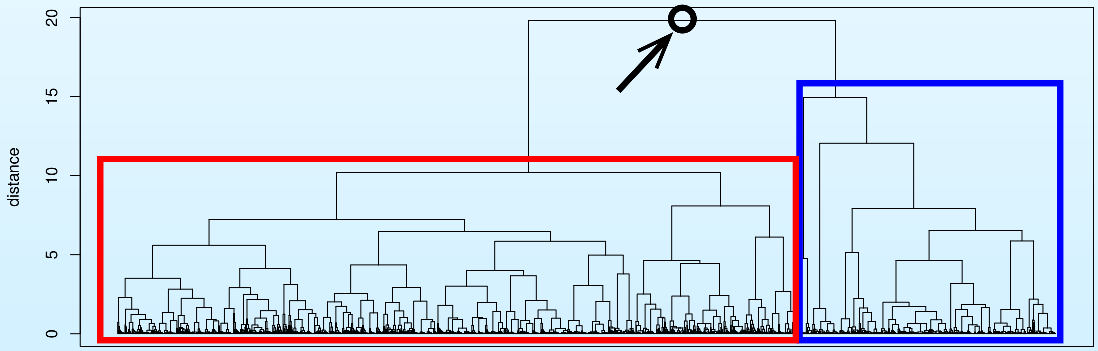
- every split in the tree defines two groups



- use Linear Discriminant Analysis to rank the acoustic features with respect to that grouping

Results

- **First split: vowels / consonants + silence**

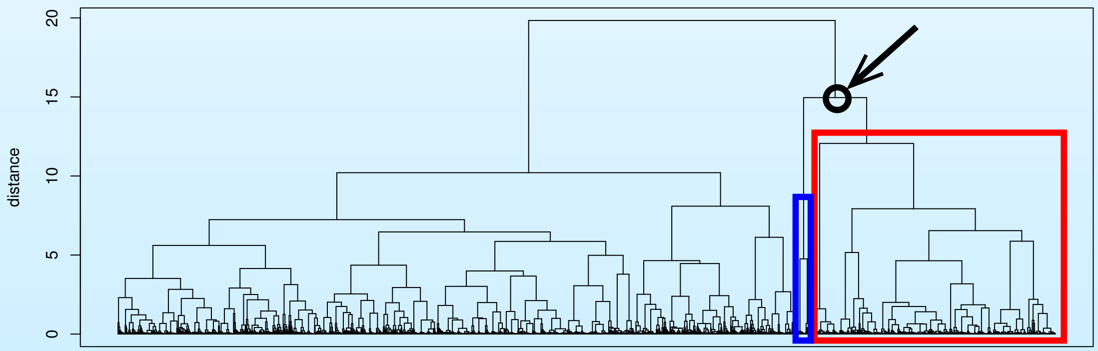


- **Discriminant analysis:**

features	prediction accuracy
c_0	78.6%
c_0, d_0	90.6%
c_0, d_0, c_2	91.4%
. . .	
all	99.5%

Results

- **Second split: silence (initial,final) / consonants + silence (middle)**

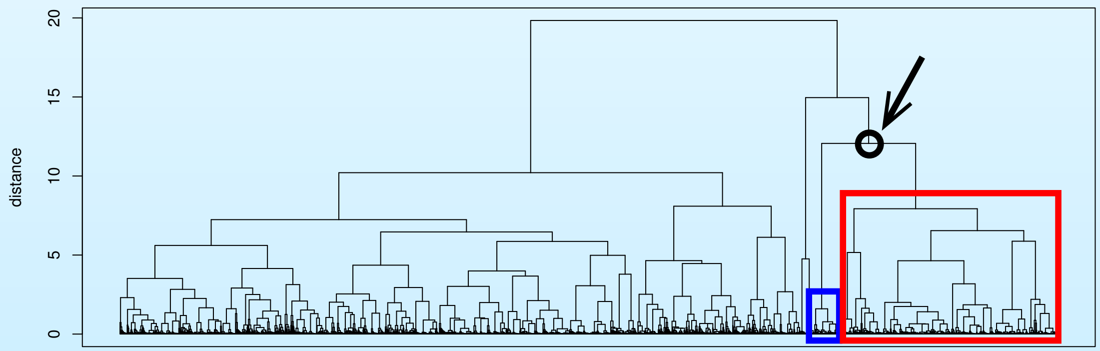


- **Discriminant analysis:**

features	prediction accuracy
a_0	94.7%
. . .	
all	100%

Results

- **Third split: voiced plosives / consonants + silence (middle)**

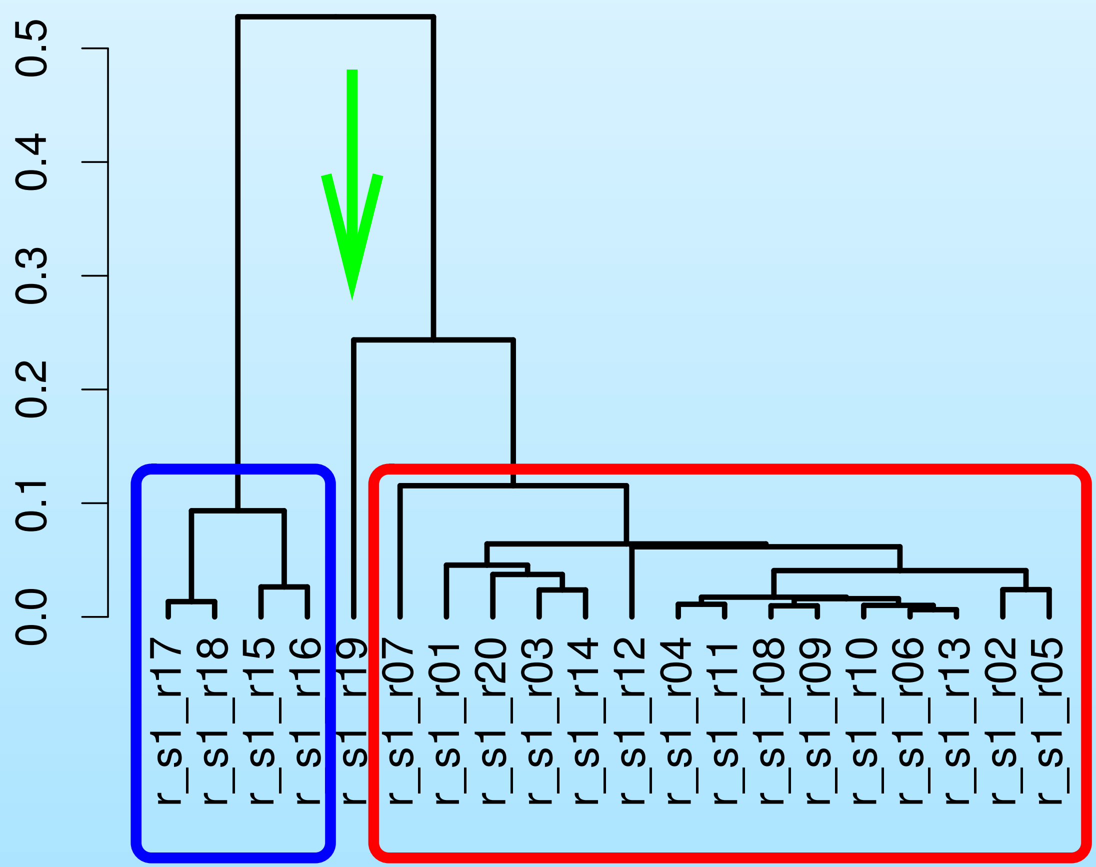


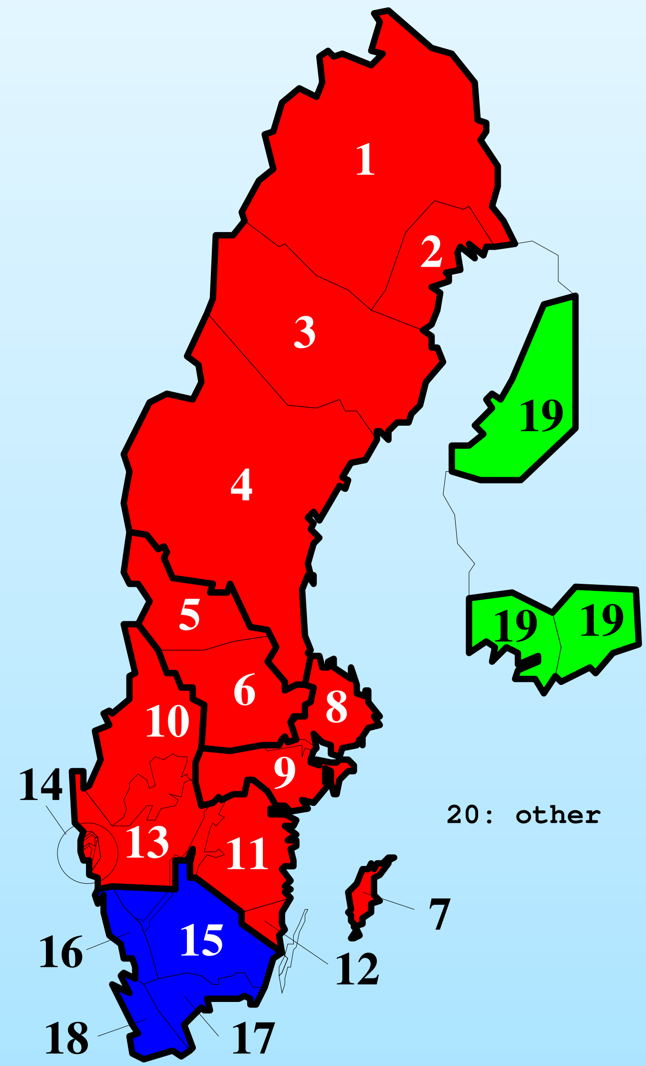
- **Discriminant analysis:**

features	prediction accuracy
d_0	88.8%
d_0, d_1	91.7%
d_0, d_1, a_9	100%
. . .	
all	100%

Results

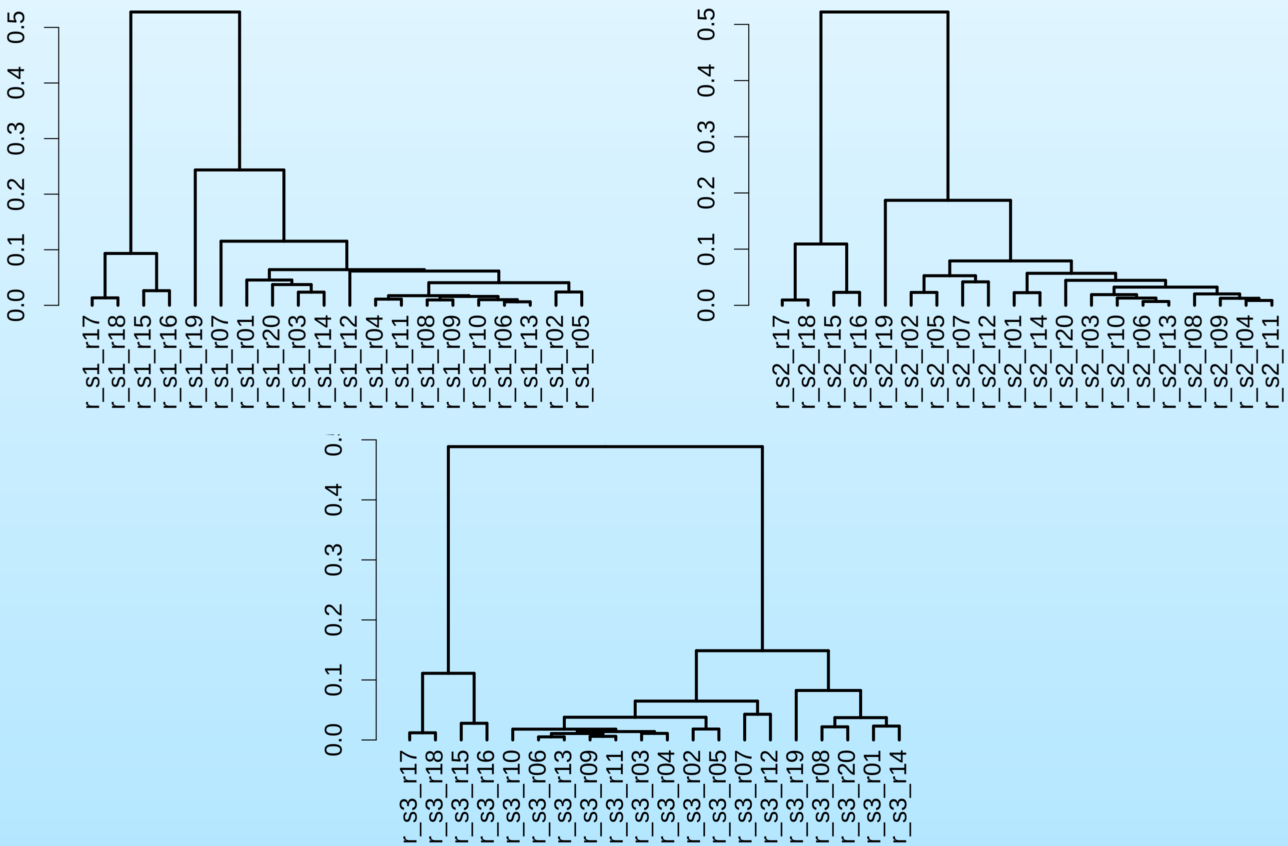
- Phoneme /r/ has a retracted pronunciation south of Sweden





Results /r/

- **Similar behaviour for initial, middle and final segment**



- **LDA: many variables explain, e.g.** c_4, d_4

Conclusions

- **The method proposed enables:**
  - analysis of large amounts of data
  - formalisation of the experiments (reproducibility)
  - analysis of cross-phoneme allophone clusters
  - separation of subsegments (initial, middle and final)
  - analysis of both broad and detailed classes of phonemes
  - ranking of the acoustic features relevant to a discrimination

Conclusions

- The method proposed enables:
  - analysis of large amounts of data
  - formalisation of the experiments (reproducibility)
  - analysis of cross-phoneme allophone clusters
  - separation of subsegments (initial, middle and final)
  - analysis of both broad and detailed classes of phonemes
  - ranking of the acoustic features relevant to a discrimination
- To do
  - interpret the results (!)
  - repeat analysis without energy features

Bibliography

http://www.speech.kth.se/~giampi

Minematsu, N. and Nakagawa, S. (2000). Visualization of pronunciation habits based upon abstract representation of acoustic observations. In *InSTIL'2000*, pages 130–137.

Salvi, G. (2003a). Accent clustering in Swedish using the Bhattacharyya distance. In *15th ICPhS Internamtional Congress of Phonetic Sciences*.

Salvi, G. (2003b). Using accent information in ASR models for Swedish. In *Eurospeech, 8th European conference on speech communication and technology*, pages 2677–2680.