# Using Accent Information in ASR Models for Swedish


GIAMPIERO SALVI

Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden

giampi@speech.kth.se

KTH Speech, Music and Hearing

Centre for Speech Technology



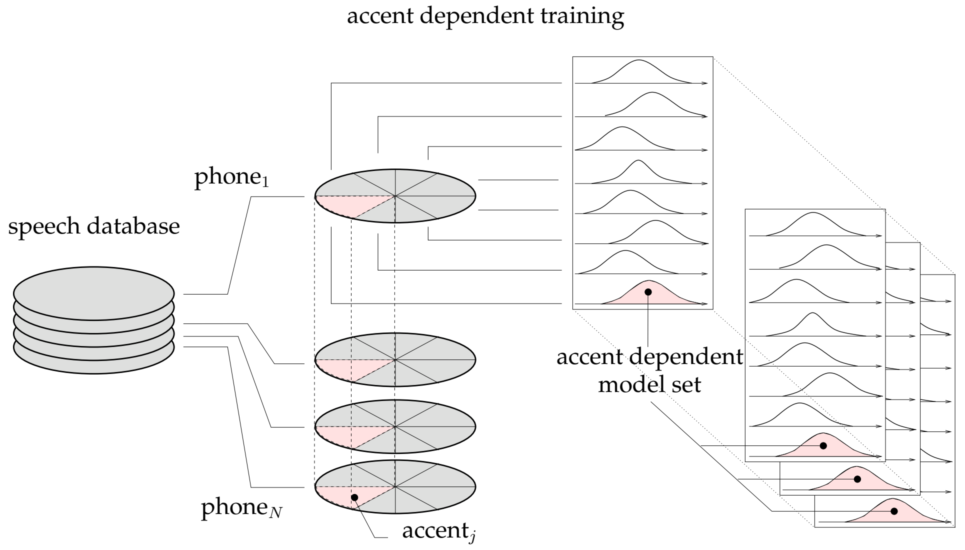
## ABSTRACT

A common technique to cope with the large variability in the acoustic realisations of the phonetic classes in speech, is to partition the data according to a linguistically significant variable. In this work, accent dependent phonetic models were trained and used both as an **analysis tool for pronunciation variation** and in the attempt to **improve ASR performance**.


## The Idea

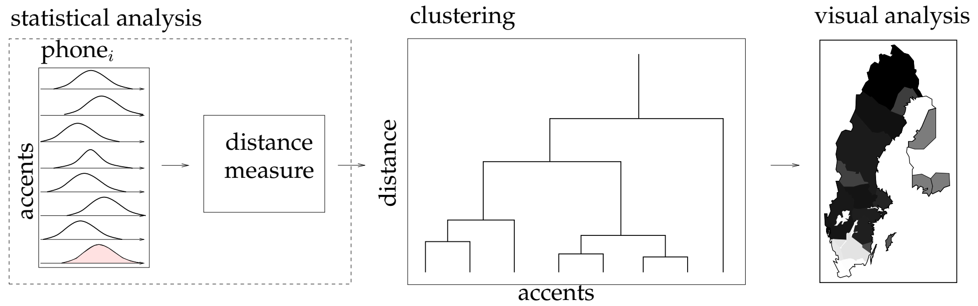
### Accent dependent training

The database is partitioned into accent areas. **Accent dependent phonetic models** are trained independently.



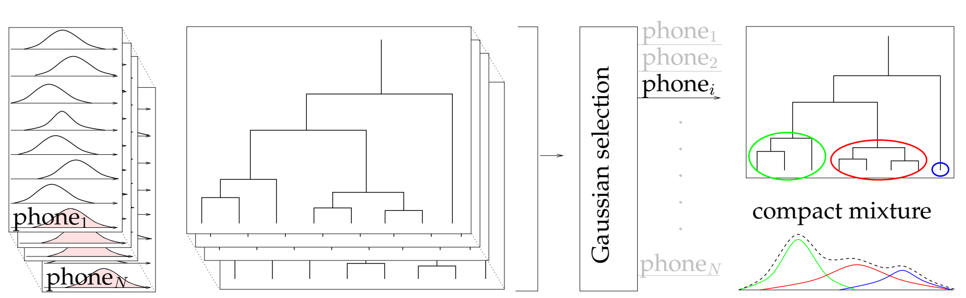

### Accent analysis ★

The model parameters obtained this way, represent the **statistical variation** of the acoustic features across **accent areas**. This information can be used for **pronunciation variation analysis**.




### Gaussian selection and ASR ▲

The distance measure in conjunction with clustering techniques can be used to select the **most representative distributions** to be assigned to each phoneme in a new ASR model set.




## The Details

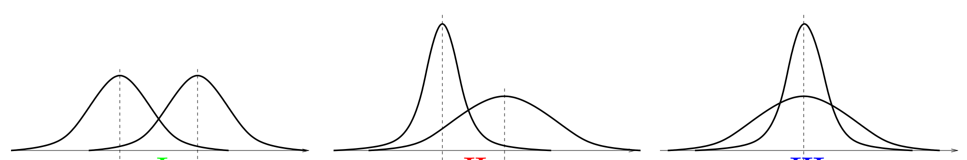
### Training

**feature extraction:** 13 MFCCs + $\Delta$ and $\Delta\Delta$.
**accent dependent phonetic models:** three states, single Gaussian, context independent.
**training data:** Swedish SpeechDat, 5000 speakers.

### Clustering

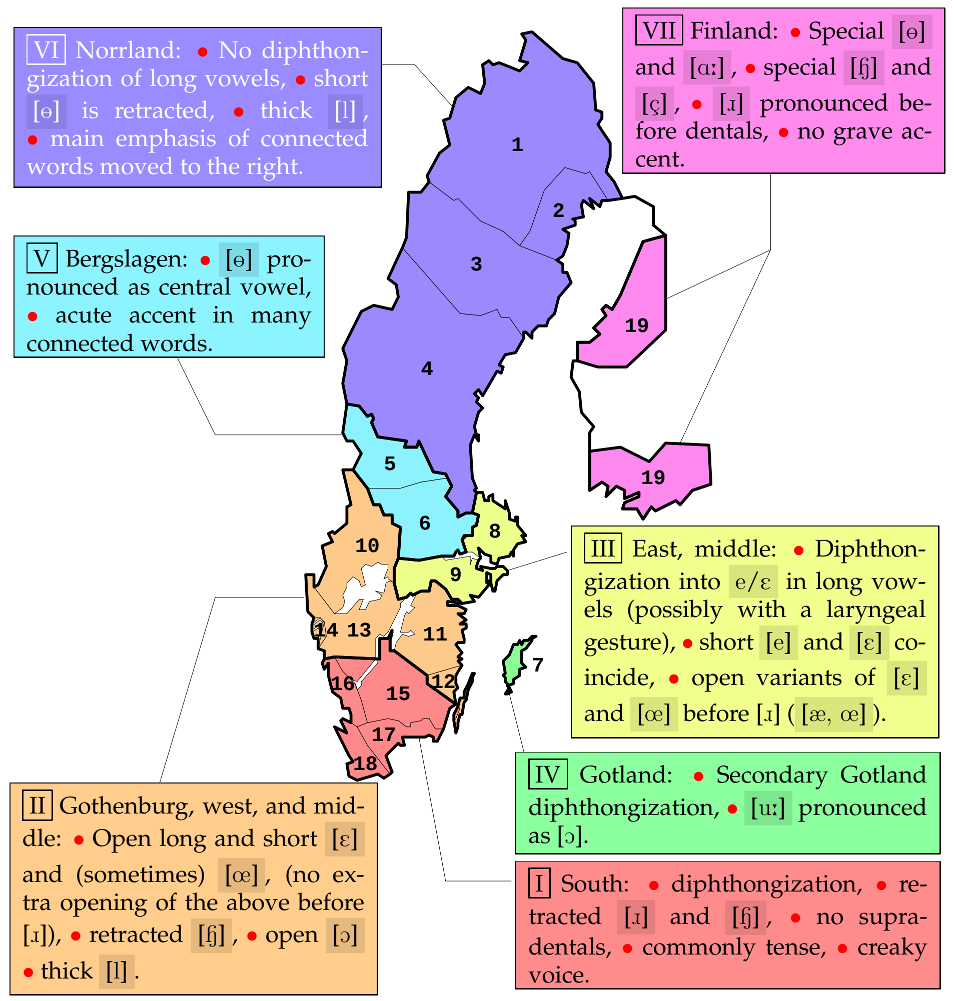
**method:** hierarchical agglomerative, complete linkage.
**metric:** Bhattacharyya distance:

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




### Accent variations in Swedish

The SpeechDat database is divided into **7 major**, and **20 minor accent areas**.

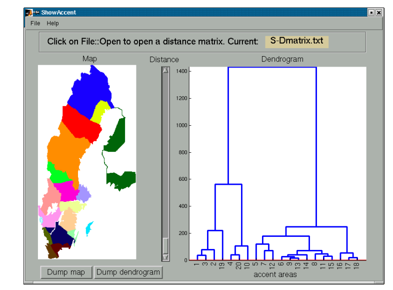



## Accent analysis ★

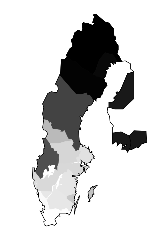
### Visual representations

**The clustering tree** The clustering tree (or **dendrogram**) is a **compact and complete representation** of the history of clustering, but it is **hard to interpret** in terms of accent areas.

**The interactive map** A tool was developed that links a map to the dendrogram. The user **selects a distance level** and the tool displays the corresponding **clusters on the map**. More intuitive, but requires selecting a distance level.



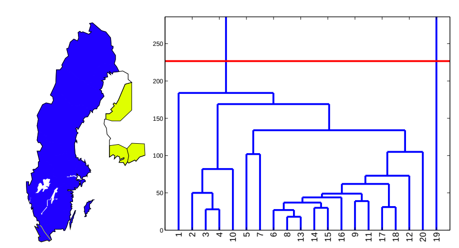
**Distance based maps** One way to preserve **both advantages** of the above representations is to use a continuous range of colors (gray levels), so that areas that are similar in pronunciation are represented by similar colors. A method was developed for this purpose, but, for simplicity, the maps are not shown in the following examples.



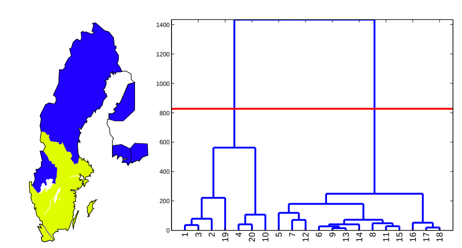
[ɧ]

### Analysis results

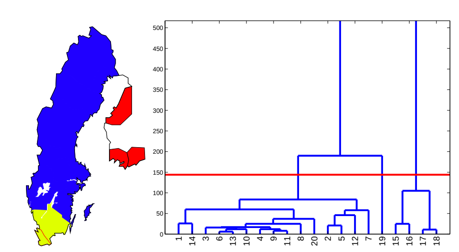
The figures show the interactive map, and the dendrogram (with distance level used in the map), for a few phonemes.



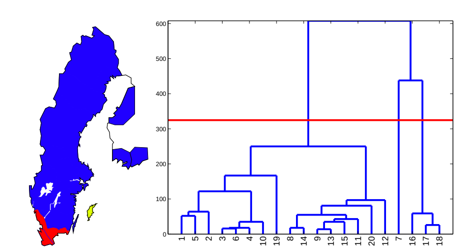
The fricative [ç] is pronounced homogeneously in Sweden, but it is an affricate in Finland.



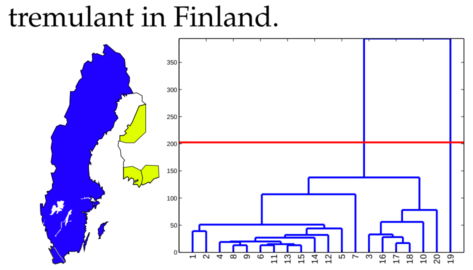
The fricative [ɧ] is more retracted in the south and middle Sweden than in the north.



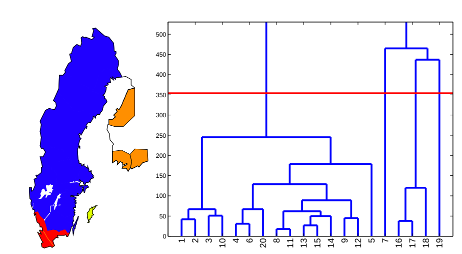
The "r" sound is an alveolar approximant [ɹ] in most of Sweden, but is retracted in the south and more tremulant in Finland.



In Gotland [ʉː] is pronounced as [ɔ], in the south of Sweden it is a diphthong.



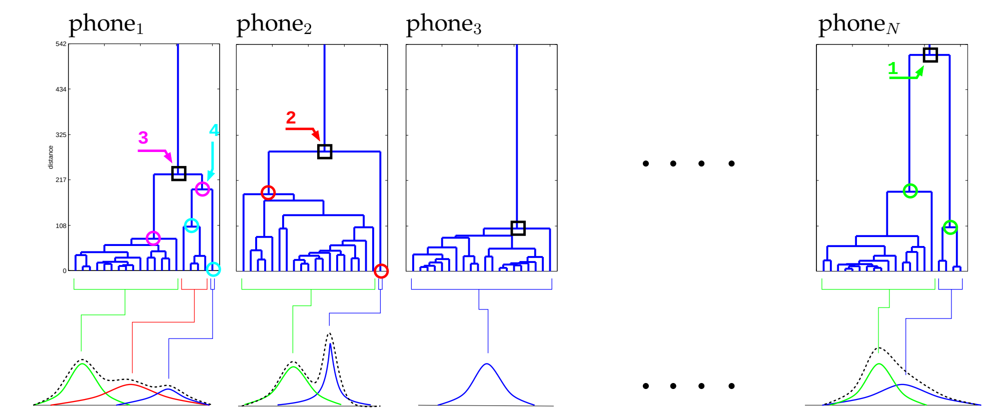
In Finland [ɑː] is pronunced as [aː]



The pronunciation of [ʉː] varies in Gotland, south and Finland.

## Gaussian selection and ASR ▲

Initially each phone is represented by one cluster (□ in the figure). At each iteration, the cluster with the highest distance level is split into the corresponding sub-clusters (e.g. ○, ○, ○, ○). When the desired total number of clusters is reached, one Gaussian component is chosen to represent each of the resulting clusters. As the figure shows, **phonemes with larger pronunciation variation** are represented by a **higher number of Gaussian components**. Eventually the model set can be retrained.




### Recognition results

Phonetic models obtained with Gaussian selection from accent dependent models (GSADs) were compared to standard Gaussian mixture models (GMMs) with the same number of components, on an isolated word task. Preliminary results show that the GSADs are slightly superior to the GMMs for low number of Gaussian components (300), while they are no better, or worse when the total number of components is higher.

## Conclusions

ASR training techniques and Bhattacharyya distance based clustering are **powerful tools for pronunciation variation analysis**.

Possible **improvements** include **releasing the assumption** that different pronunciations for each phoneme can **only merge within the corresponding phonetic class**.

The use of accent information in **ARS models** is promising, but needs **further refinements**:

- better **cluster selection** algorithm
- better choice for the **distribution** that is to represent each cluster after selection.
- include also **more prominent sources of variablility** (e.g. gender)