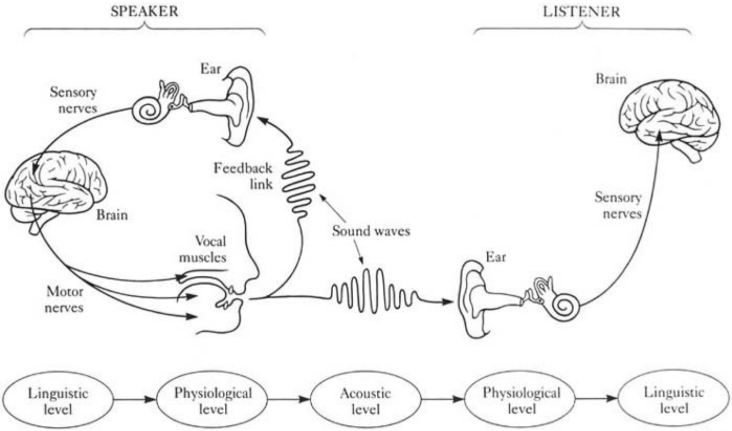
Ecological Language Acquisition via Incremental Model-Based Clustering

Giampiero Salvi

KTH CSC TMH giampi@kth.se

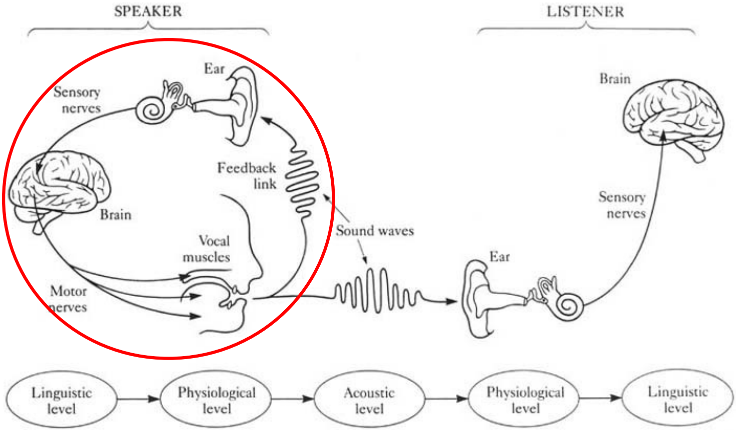
Nov. 2005

The Speech Chain



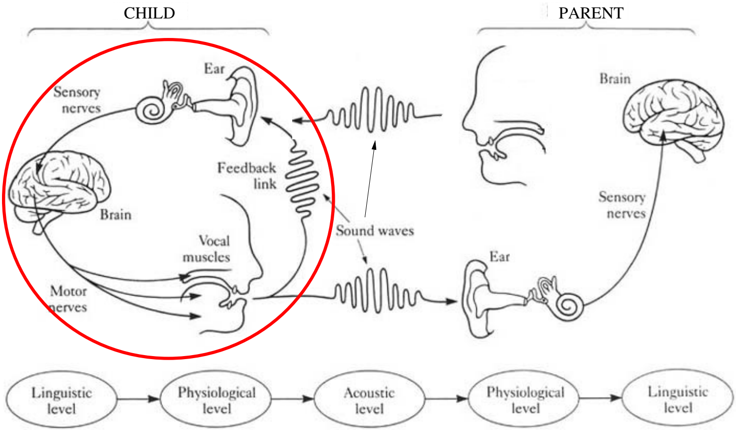
Denes and Pinson (1993)

The Speech Chain



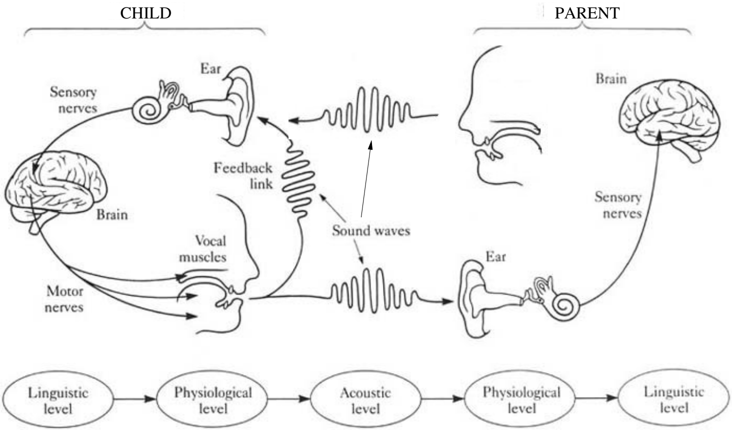
Denes and Pinson (1993)

The Speech Chain



Denes and Pinson (1993)

The Speech Chain



Denes and Pinson (1993)

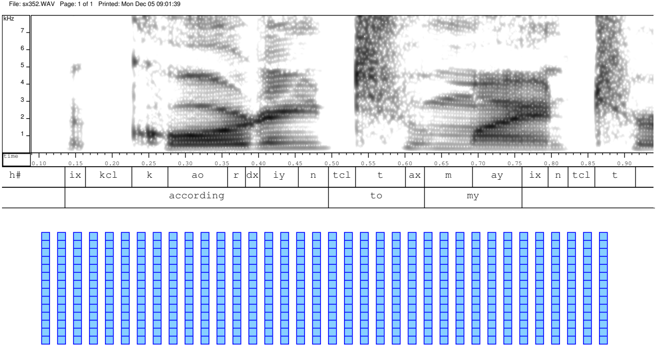
- Background: ecological theory of language acquisition (Lacerda et al., 2004)
  - the infant is naïve: no innate linguistic knowledge

- Background: ecological theory of language acquisition (Lacerda et al., 2004)
  - the infant is naïve: no innate linguistic knowledge
- Aim (long term): mathematical modelling of the learning process
  - acoustic features classification
  - time integration into meaningful sequences
  - integration of acoustic/visual information

- Background: ecological theory of language acquisition (Lacerda et al., 2004)
  - the infant is naïve: no innate linguistic knowledge
- Aim (long term): mathematical modelling of the learning process
  - acoustic features classification
  - time integration into meaningful sequences
  - integration of acoustic/visual information
- Aim Interspeech 2005 (Salvi, 2005): acoustic features classification
  - unsupervised
  - incremental

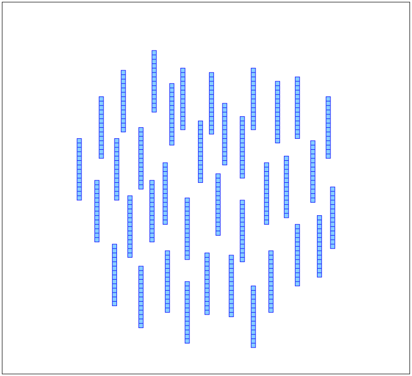
Acoustic features

Equally spaced windows of speech



Assumption

Acoustic feature vectors independently drawn from mixture of Gaussian distributions



Method

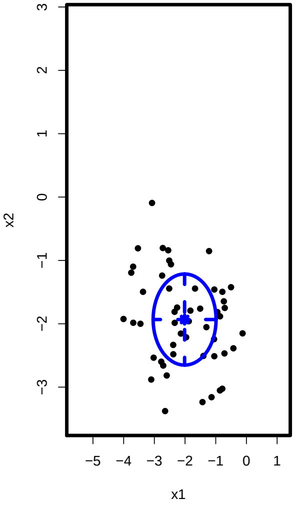
- Model-Based Clustering (Fraley and Raftery, 1998)
  - data modelled as mixture of probability distributions
  - each distribution represents a cluster
  - each data point belongs to each cluster with a certain probability
  - model parameters estimated via Expectation Maximisation
  - different models compared via Bayes information criterion (BIC)

Method

- Model-Based Clustering (Fraley and Raftery, 1998)
  - data modelled as mixture of probability distributions
  - each distribution represents a cluster
  - each data point belongs to each cluster with a certain probability
  - model parameters estimated via Expectation Maximisation
  - different models compared via Bayes information criterion (BIC)
- Incremental Model-Based Clustering (Fraley et al., 2003)
  - introduced for large datasets

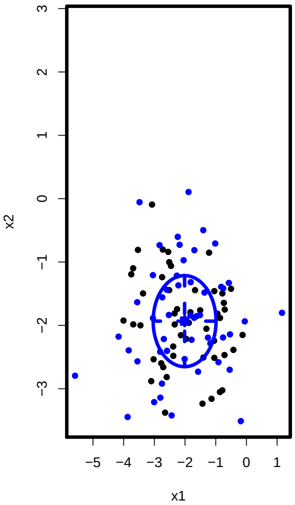
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



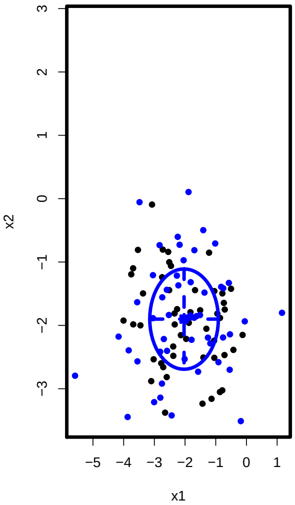
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



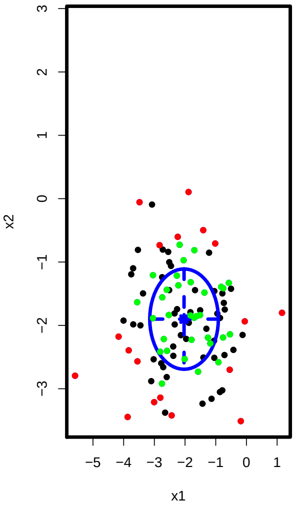
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



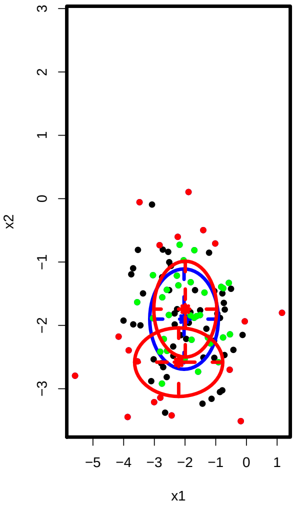
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



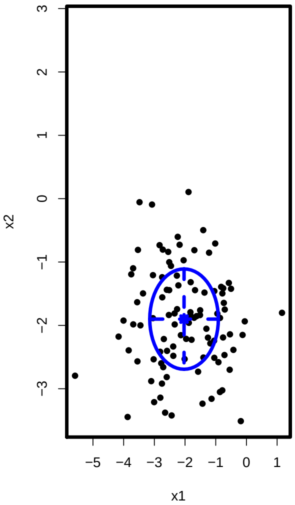
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



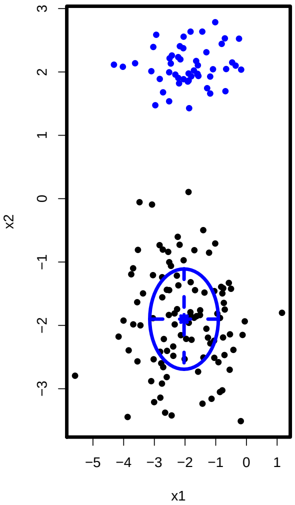
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



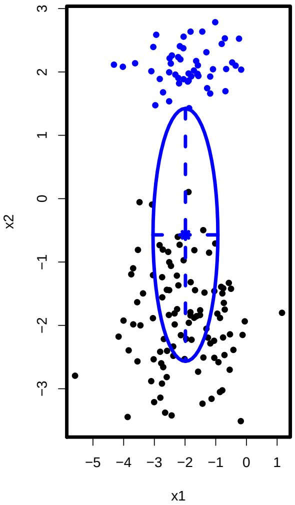
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



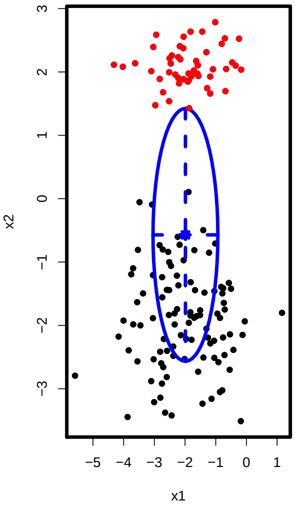
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



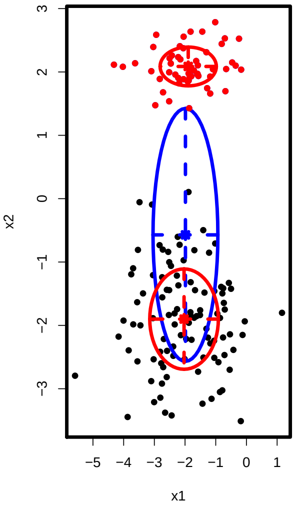
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



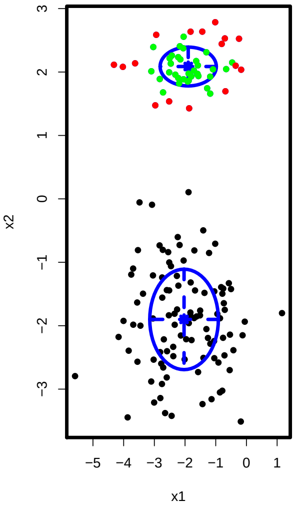
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



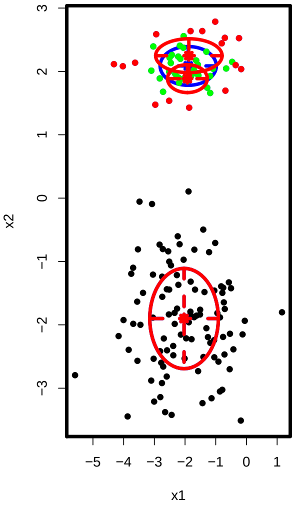
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



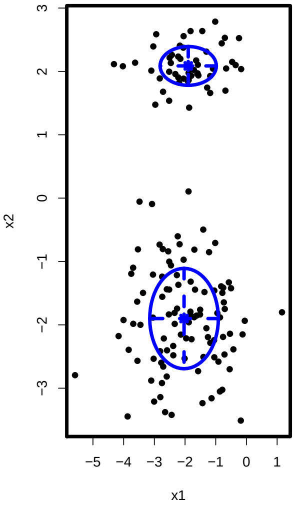
Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



Algorithm

1. start with a MCLUST model
2. get new data
3. adjust old model to new data
4. divide new data into well and poorly modelled points
5. try a more complex model, if better BIC set as best and go back to 4
6. set the current best model and go back to 2



Experimental settings

- Data (ex1, ex2, ex3, ex4, ex5)
  - 12 minutes from the MILLE corpus
  - child directed speech (1 mother talking to her child)
  - Mel frequency cepstral coeffs computed every 10ms + differences of first and second order

Experimental settings

- Data (ex1, ex2, ex3, ex4, ex5)
  - 12 minutes from the MILLE corpus
  - child directed speech (1 mother talking to her child)
  - Mel frequency cepstral coeffs computed every 10ms + differences of first and second order
- experimental factors
  - dimensionality of the data: from 3 to 39 dimensions
  - frame length: from 200msec to 3sec

Evaluation

- problem: there is no reference (at the moment)

Evaluation

- problem: there is no reference (at the moment)
- relative evaluation:

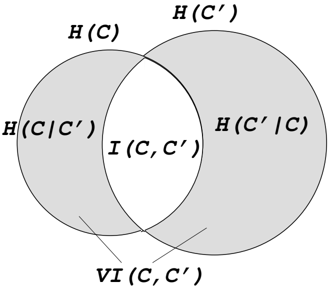
Evaluation

- problem: there is no reference (at the moment)
- relative evaluation:
- time evolution of number of clusters
  - dependency with number of feature coefficients
  - dependency with frame length

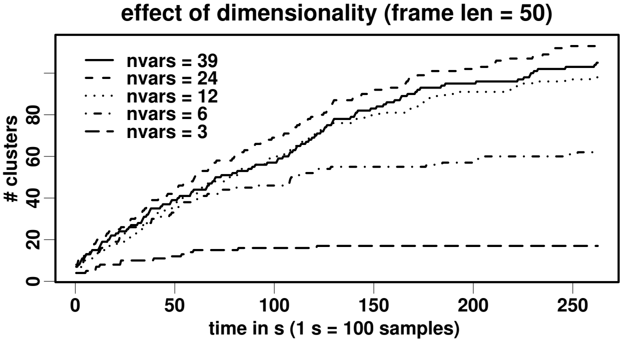
Evaluation

- problem: there is no reference (at the moment)
- relative evaluation:
- time evolution of number of clusters
  - dependency with number of feature coefficients
  - dependency with frame length
- agreement of classification in different conditions
  - variation of information (Meilă, 2002)

$$\mathbf{V}I(C, C') = H(C|C') + H(C'|C)$$

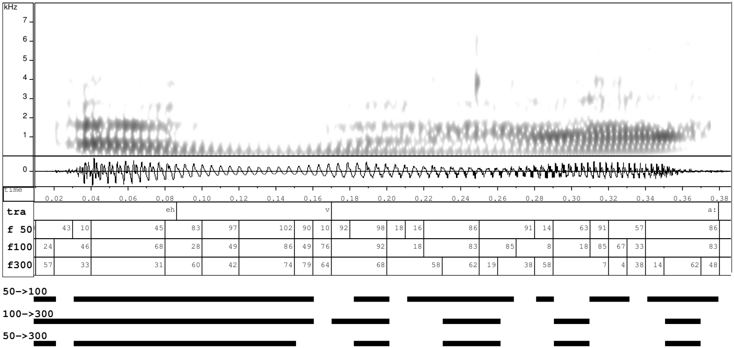


Results



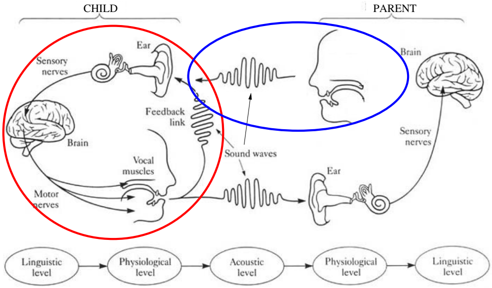
effect of dimensionality (frame len = 50)
nvars = 39
nvars = 24
nvars = 12
nvars = 6
nvars = 3
# clusters
0
20
40
60
80
0
50
100
150
200
250
time in s (1 s = 100 samples)

Example

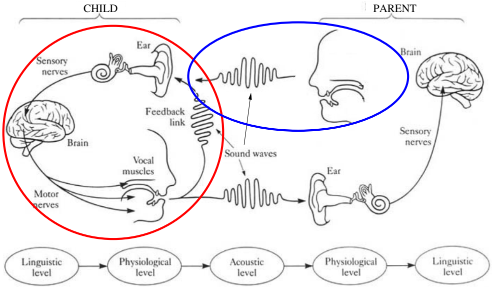


kHz
7
6
5
4
3
2
1
0
time
0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.38
tra eh v a:
f 50 43 10 45 83 97 102 90 10 92 98 18 16 86 91 14 63 91 57 86
f100 24 46 68 28 49 86 49 76 92 18 83 85 8 18 85 67 33 83
f300 57 33 31 60 42 74 79 64 68 58 62 19 38 58 7 4 38 14 62 48
50->100
100->300
50->300

Mismatch Child/Parent Voice

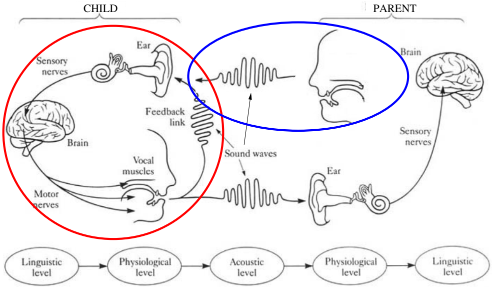


Mismatch Child/Parent Voice



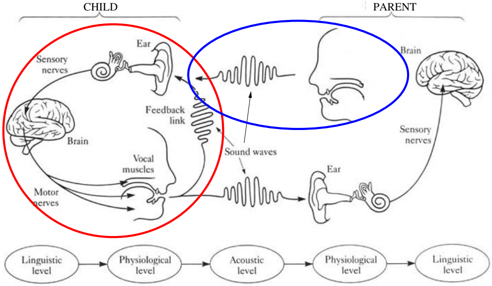
- ASR with children

Mismatch Child/Parent Voice



- ASR with children
- Normalisation
  - VTLN: Vocal Tract Length Normalisation
  - Adaptation: hard in this context

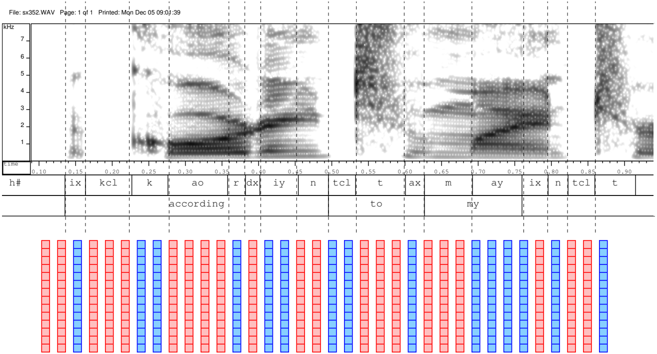
Mismatch Child/Parent Voice



- ASR with children
- Normalisation
  - VTLN: Vocal Tract Length Normalisation
  - Adaptation: hard in this context
- Relative Features

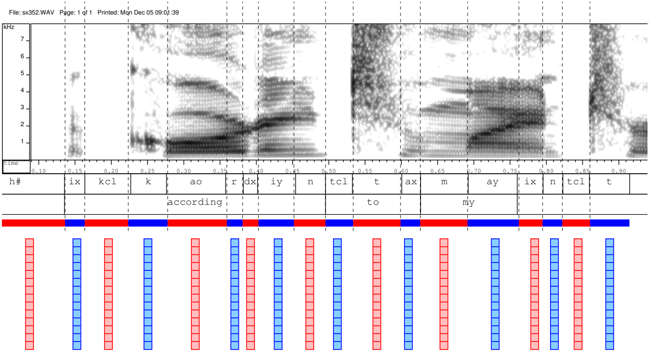
Acoustic Features

Frame Based



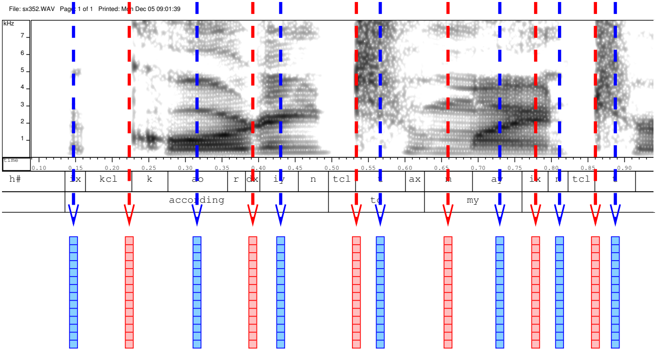
Acoustic Features

Segment Based

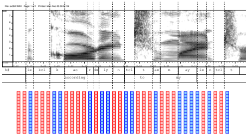


Acoustic Features

Landmark Based

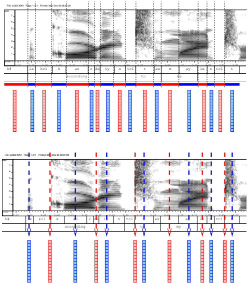


Consequences



⇒

Sequence recognition (HMMs)

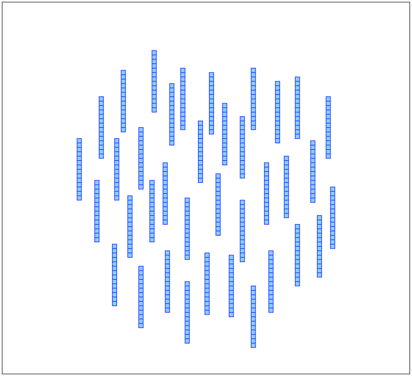


⇒

simpler relation acoustic categories/ linguistic units

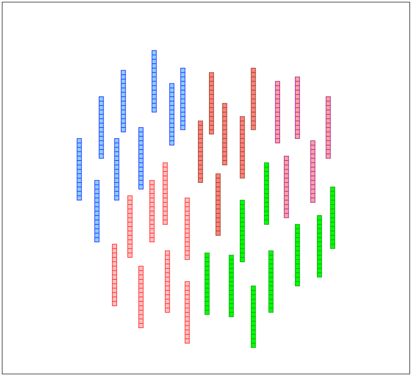
Clustering Time Sequences

Acoustic vectors independently drawn from mixture of gaussian distributions



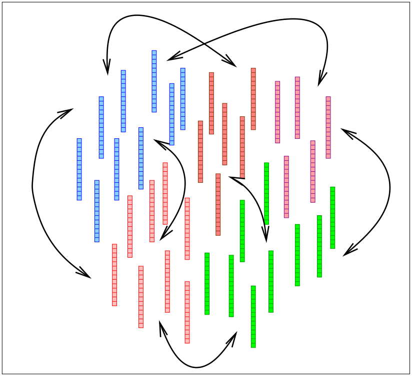
Clustering Time Sequences

Acoustic vectors independently drawn from mixture of gaussian distributions

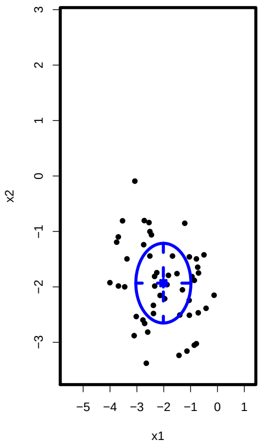


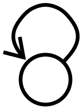
Clustering Time Sequences

Acoustic vectors independently drawn from mixture of gaussian distributions

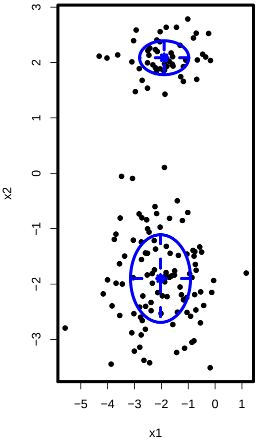


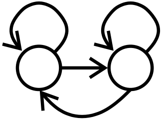
Modeling time evolution with Markov chains



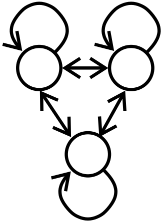


Modeling time evolution with Markov chains

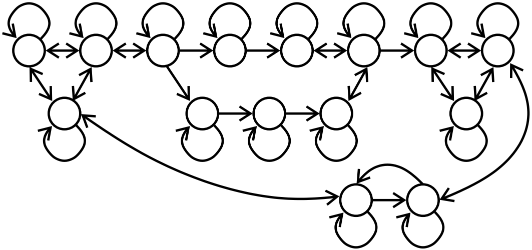




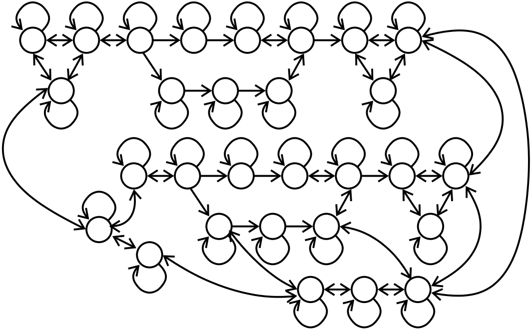
Modeling time evolution with Markov chains



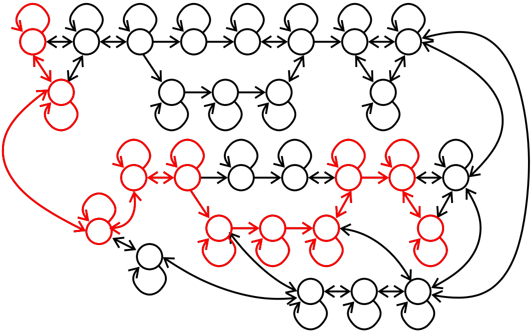
Modeling time evolution with Markov chains



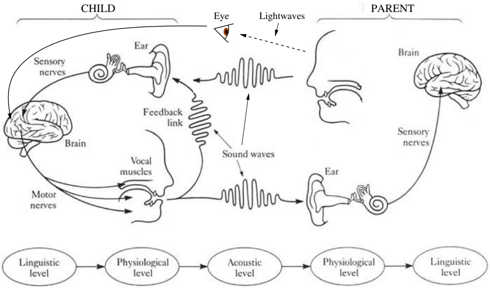
Modeling time evolution with Markov chains



Modeling time evolution with Markov chains

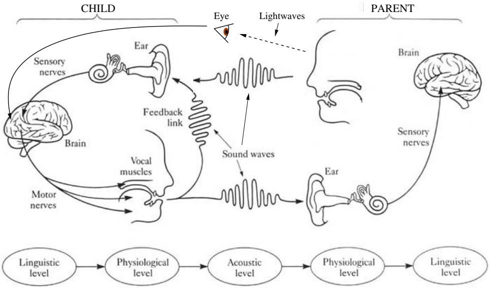


The Visual Channel



- No one-to-one relation acoustic/visual info

The Visual Channel



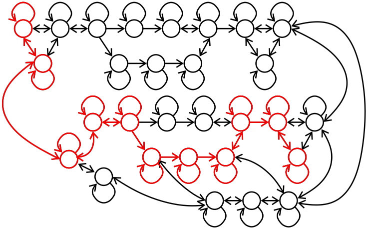
- No one-to-one relation acoustic/visual info
- Reinforcement Learning
  - perform match at higher levels (pseudo-words or -phrases)

The Visual Channel

Perform visual/acoustic match on the Markov chain

Acoustic Event

Visual Event

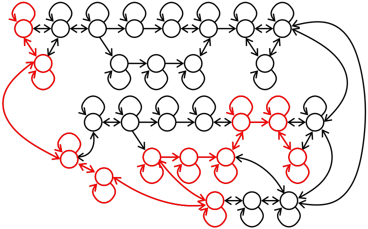


The Visual Channel

Perform visual/acoustic match on the Markov chain

Acoustic Event

Visual Event

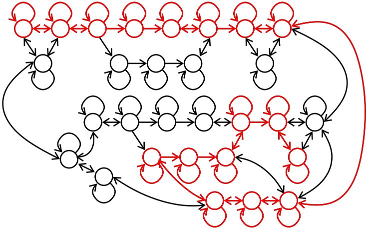


The Visual Channel

Perform visual/acoustic match on the Markov chain

Acoustic Event

Visual Event

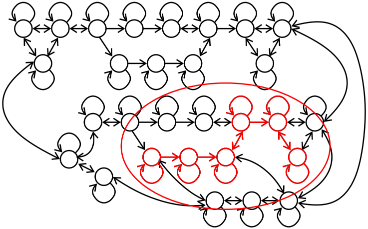


The Visual Channel

Perform visual/acoustic match on the Markov chain

Visual Event

Acoustic Event



The Final Question

- Are the acoustic blocks (categories) in a language learned out of their statistical occurrence or out of their contrastive use?

The Final Question

- Are the acoustic blocks (categories) in a language learned out of their statistical occurrence or out of their contrastive use?
- in the first case: model based clustering and growing Markov chains are separate processes.

The Final Question

- Are the acoustic blocks (categories) in a language learned out of their statistical occurrence or out of their contrastive use?
- in the first case: model based clustering and growing Markov chains are separate processes.
- in the second case: need to integrate everything

Bibliography

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

Denes, P. B. and Pinson, E. N. (1993). *The Speech Chain: Physics and Biology of Spoken Language*. W. H. Freeman.

Fraley, C., Raftery, A., and Wehrensy, R. (2003). Incremental model-based clustering for large datasets with small clusters. Technical Report 439, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *Computer Journal*, 41(8).

Lacerda, F., Klintfors, E., Gustavsson, L., Lagerkvist, L., Marklund, E., and Sundberg, U. (2004). Ecological theory of language acquisition. In *EPIROB*, pages 147–148.

Meilă, M. (2002). Comparing clusterings. Technical Report 418, Department of Statistics, University of Washington.

Salvi, G. (2005). Ecological language acquisition via incremental model-based clustering. In *Interspeech*, pages 1181–1184.